DOCUMENT RESUME

ED 260 083 TM 850 168

AUTHOR Norris, Stephen P.; King, Ruth
TITLE The Design of a Critical Thinking Test on Appraising Observations. Studies in Critical Thinking, Research Report No. 1.
INSTITUTION Memorial Univ., St. John's (Newfoundland). Inst. for Educational Research and Development.
SPONS AGENCY Illinois Univ., Urbana. Bureau of Educational Research.; Social Sciences and Humanities Research Council of Canada, Ottawa (Ontario).
PUB DATE 84
GRANT 418-81-0781
NOTE 148p.
PUB TYPE Tests/Evaluation Instruments (160) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS *Cognitive Tests; *Critical Thinking; Diagnostic Tests; Difficulty Level; Educational Diagnosis; High Schools; Item Analysis; Meta Cognition; *Observation; Observational Learning; Research Methodology; Self Evaluation (Individuals); Student Attitudes; *Test Construction; Test Interpretation; *Test Validity
IDENTIFIERS Protocol Analysis; *Test on Appraising Observations

ABSTRACT

This report describes the design of a test of one aspect of critical thinking ability, the ability to correctly appraise observations. Intended for classroom use with senior high school students, the 50 item Test on Appraising Observations is based on a comprehensive set of principles modified from Robert Ennes' conception of good observation appraisal. The test evolved through many versions using methodology developed from the construct validity theory that ability tests are valid to the extent that good thinking leads to good test performance and that poor thinking leads to bad test performance. The methodology involved the systematic collection of thinking about protocols of examinees while they worked through test questions. Data collection and analysis for two experimental test versions are described in detail. Item and test statistics for the final version were collected from four southern Ontario high schools and compared to results of two other critical thinking tests. The appendices contain two versions of the Test on Appraising Observations; the Test on Assessing the Believability of Observation Statements; the Observation Test Interview Model, B; the Instruction Sheet to Cooperating Teachers; and the answer key and principles tested per item. (BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED260083

The Design of a Critical Thinking Test on Appraising Observations

Stephen P. Norris, Principal Investigator

and

Ruth King, Research Associate

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
S. P. Norris
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Studies in Critical Thinking

Research Report No. 1

Institute for Educational Research and Development

Memorial University of Newfoundland

St. John's, Newfoundland

Canada

A1B 3X8

Copyright © 1984 Stephen P. Norris

This research was supported in part by the Bureau of Educational Research, University of Illinois at Urbana-Champaign while I was a doctoral candidate there, and in part by a grant from the Social Sciences and Humanities Research Council of Canada (Grant No.: 418-81-0781), and in part by a grant from the Vice-president, Memorial University of Newfoundland.

TM 850 168

BEST COPY

Table of Contents

1. Introduction 1

1.1. The Perceived Need for Critical Thinking Research 2

2. Critical Thinking Testing and Conceptualization 5

2.1. Critical Thinking Tests 6

2.2. Critical Thinking Conceptualization 9

3. Principles for Appraising Observations 13

3.1. Justifying the Principles 16

4. Test Validation 21

4.1. The Basis of a Theory of Construct Validation 21

4.1.1. Theoretical Terms 22

4.1.2. Causal Explanation 24

4.1.3. Standards of Adequacy for Judging Theories 26

4.2. Assumptions of the Validation Methodology Used in this Study 27

5. The Evolution of the Test on Appraising Observations 31

5.1. Decisions Concerning Audience and Style 31

5.1.1. Audience 31

5.1.2. Style 32

5.2. Interaction of Validation Methodology and Stages of Development 35

5.3. Transitions Between Test Versions 40

5.3.1. Preliminary Versions 40

5.3.2. Experimental versions 42

5.4. Summary 45

6. Data Collection and Analysis: Final Results 47

6.1. Samples and Data Collection 47

6.2. Basic Data and Derived Data 49

6.2.1. Basic Data 49

6.2.2. Derived Data 50

6.3. Analysis: Version B to Version C 55

6.3.1. Thinking/Performance Relationships for Items 56

6.3.2. Thinking/Performance Relationship for the Test as a Whole 60

6.3.3. Items Testing the Same Principle 61

BEST COPY

6.3.4. The Relationship of Item Performance to Overall Test Performance 63
6.3.5. Test Wiseness 63
6.3.6. Misleading Factors, Reading Difficulty, and Clarity of Instructions 66
6.3.7. Test Characteristics and Student Ability 67
6.3.8. Extraneous Influences on Performance Scores and Thinking Scores 69
6.3.9. Summary 71
6.4. Analysis: Version C 72
6.4.1. Thinking/Performance Relationships for Items 72
6.4.2. Thinking/Performance Relationship for the Test as a Whole 75
6.4.3. Items Testing the Same Principle 75
6.4.4. Extraneous Influences on Performance Scores and Thinking Scores 76
6.5. Final Data 80
7. Summary and Conclusions 85
Appendix A. Test on Appraising Observations 89
Appendix B. Test on Assessing the Believability of Observation Statements 101
Appendix C. Observation Test Interview Model, B 117
C.1. Interview Sheets 122
Appendix D. Test on Appraising Observations, Version C 127
Appendix E. Instruction Sheet to Cooperating Teachers 147
Appendix F. Key to Correct Answers and Principles Tested per Item on the Test on Appraising Observations 149

List of Figures

Figure 6-1: Scatterplot of Test Performance versus 68 Thinking/Performance Biserial Correlation

List of Tables

Table 3-1:	Principles for Appraising Observations	14
Table 6-1:	Distribution of Samples by Grade Level, School, and Testing Format	48
Table 6-2:	Number of Students in Ontario Samples for TAO, CCTT, and W-G	49
Table 6-3:	Rating Scale for Thinking Scores	51
Table 6-4:	Rating Thinking for Item 3	52
Table 6-5:	Weights of Evidence for Thinking/Performance Combinations	54
Table 6-6:	Item Level Statistics, Version B	57
Table 6-7:	Pearson Correlations Between Thinking Scores and Performance Scores	60
Table 6-8:	Item Difficulties and Item/Test Biserial Correlations for Items Testing the Same Principles	62
Table 6-9:	Average Performance Scores by Test Section and by Order of Interviewing	66
Table 6-10:	Average Thinking Scores by Test Section	66
Table 6-11:	ANOVA Results for Version B: Performance Scores by Testing Format, Interviewer, Test Section, Grade Level, and Sex of Examinee	70
Table 6-12:	MANOVA Results for Version B: Performance Scores and Thinking Scores by Interviewer, Test Section, Grade Level, and Sex of Examinee	71
Table 6-13:	Comparision of T/P Index Scores Per Item for Version B and Version C	73
Table 6-14:	ANOVA Results for Version C: Performance Scores by Testing Format, Interviewer,Test Section, Grade Level, and Sex of Examinee	77
Table 6-15:	MANOVA Results for Version C: Performance Scores and Thinking Scores by Interviewer, Test Section, Grade Level, and Sex of Examinee	78
Table 6-16:	Item Analysis Results for All Southern Ontario Schools Combined	80
Table 6-17:	Item Analysis Results for Bluevale School	81
Table 6-18:	Item Analysis Results for Forest Heights School	81

Table 6-19: Item Analysis Results for Guelph School 82
Table 6-20: Item Analysis Results for Waterloo School 82

Chapter 1
Introduction

This report describes the design of a critical thinking test of high school students' ability to appraise observations. Statements of observation, as well as fundamental principles of value, are at the foundation of much of what we believe. If critical thinking is conceived as dealing with decisions about what to believe, as it is by Robert Ennis[1] and David Hitchcock,[2] then basic to such decisions will be appraisals of the confidence to place in particular statements of observation. This recognition, along with the fact that there is no readily-available test of this aspect of critical thinking ability, motivated this project.

In addition to the obvious outcome of a published test and manual, the project had two other major objectives. The first was the elucidation and defense of a set of principles upon which to base appraisals of observations. This has been accomplished and reported.[3] The second was the elaboration and the testing of a methodology of gaining test validation information by probing people's reasons for responding to test questions the way they do. The methodology is explained in this report, but the

---

[1]Ennis, R.H. A concept of critical thinking, *Harvard Educational Review, 32*, 1962, 81-111.

[2]Hitchcock, D. *Critical thinking: A guide to evaluating information*, Toronto: Methuen, 1983.

[3]Norris, S.P. Defining observational competence, *Science Education, 68*, 1984, 129-142.

BEST COPY

employment of the technique led to further unanswered questions which are currently being explored. In brief, I am attempting to determine to what extent people's verbal reports of their thinking on tests are accurate reflections of the thinking which takes place, and to what extent the manner of eliciting those reports influence the reports' contents. This study is yielding interesting results and will be reported in the near future.

1.1. The Perceived Need for Critical Thinking Research

Critical thinking has long been held in high esteem by educators and educational theorists. However, it is only recently that there has been support for a concerted effort to implement critical thinking instruction into grade school and university. The result has been the introduction of critical thinking materials (although they are not always called 'critical thinking') at all educational levels, and a growing network of communication among those interested in teaching and theorizing about critical thinking. For example, the primary and elementary grades have seen the development of a philosophy for children course, which has had increasing popularity throughout the world. One unit of this programme, *Harry Stottlemeier's Discovery*,[4] concentrates on teaching children how to think philosophically. In addition, the institute which developed this programme publishes a journal, *Thinking*, devoted to theoretical and practical problems in teaching critical thinking.

At California State University critical thinking is being taken seriously to the point where the Chancellor of the University issued an order, Executive order 338, making the study of critical thinking a requirement for every undergraduate in the system. One result of this move has been the initiation of a critical thinking newsletter, *CT News*, to facilitate communication among those interested in teaching critical

[4]Lipman, M. *Harry Stottlemeier's discovery*. Upper Montclair, New Jersey: The Institute for the Advancement of Philosophy for Children, 1977.

thinking, and the regular sponsorship of conferences on critical thinking in California.

In Canada, the current thrust in critical thinking is primarily at the university level, and involving philosophers and philosophers of education. There have been two books on the subject published in Canada in the past couple of years. One is by Hitchcock, mentioned previously, which is a text book on critical thinking methods. The other, *Critical Thinking and Education*,[5] is mainly a theoretical analysis of the concept of critical thinking and of the place of critical thinking instruction in education.

Two philosophers at the University of Windsor are responsible for the birth of a new journal devoted to critical thinking, *Informal Logic*. They have also sponsored two international conferences on critical thinking, or informal logic as it is sometimes called, which have in turn led to the institution of a new society, The Association for Informal Logic and Critical Thinking, formed to promote interest and dialogue in this area.

It seems, then, that there is a growing interest in critical thinking instruction which goes beyond the mere lip service that has often been paid to this goal of education. However, the increased interest has highlighted a number of serious deficiencies. Much work is needed on the conceptualization of what constitutes good critical thinking, on the development and testing of approaches and materials for instruction, and on designing additional techniques for assessing the extent to which people are critical thinkers. In the following chapter there is a discussion of the available critical thinking tests and of the existing analyses of critical thinking.

---

[5]McPeck, J.E. *Critical thinking and education*. Oxford: Martin Robertson, 1981.

Chapter 2

Critical Thinking Testing and Conceptualization

The way in which one conceives of critical thinking will affect how one categorizes attempts to test for it. If one thinks of critical thinking ability as a multi-faceted competence as I do and as Robert Ennis does,[6] then one tends to see two types of critical thinking tests, general ones and aspect-specific ones. General critical thinking tests are those which attempt to give an indication of people's critical thinking ability in general. That is, they test for many or all aspects of critical thinking. Aspect-specific critical thinking tests are those which attempt to test for only one aspect of critical thinking, such as inductive thinking ability, deductive thinking ability, assumption-finding ability, observation appraisal ability, or whatever the case may be. There are advantages and disadvantages to both types of tests. One gives broad coverage but virtually no detailed knowledge or diagnostic assistance. The other type gives detailed knowledge and diagnostic assistance but at the expense of a narrow focus. The test of observation appraisal designed in this study is an aspect-specific critical thinking test.

BEST COPY

[6]Ennis, R.H. A conception of rational thinking, in J.R. Coombs(Ed.), *Philosophy of Education 1979*, Normal, Illinois: The Philosophy of Education Society, 1980.

2.1. Critical Thinking Tests

At present there are five, general, easily scoreable, English language critical thinking tests readily available in North America. These are the *Cornell Critical Thinking Test, Level X*,[7] the *Cornell Critical Thinking Test, Level Z*,[8] the *New Jersey Test of Reasoning Skills*,[9] the *Ross Test of Higher Cognitive Processes*,[10] and the *Watson-Glaser Critical Thinking Appraisal*.[11] These tests purport to provide an easily usable way of acquiring a broad picture of people's critical thinking competence, although each of the tests suffers some flaws.[12]

The situation concerning aspect-specific tests of critical thinking is in far worse shape. In an extensive review of the literature on testing critical thinking Bruce Stewart[13] discovered several unpublished general critical thinking tests, and also several tests of deductive thinking ability, one aspect of critical thinking ability. However, he found few tests of other aspects of critical thinking, a glaring deficiency if one educational goal is to acquire detailed diagnostic information of students' critical thinking ability.

---

[7]Ennis, R.H. and Millman, J. *Cornell Critical Thinking Test, Level X*. Champaign, Illinois: The Illinois Thinking Project, 1982a.

[8]Ennis, R.H. and Millman, J. *Cornell Critical Thinking Test, Level Z*. Champaign, Illinois: The Illinois Thinking Project, 1982b.

[9]Shipman, V. *New Jersey Test of Reasoning Skills*. Upper Montclair, New Jersey: Institute for the Advancement of Philosophy for Children, 1983.

[10]Ross, J.D. and Ross, C.M. *Ross Test of Higher Cognitive Processes*. Novato, California: Academic Therapy Publications, 1976.

[11]Watson, G. and Glaser, E.M. *Watson-Glaser Critical Thinking Appraisal*. New York: The Psychological Corporation, 1980.

[12]Ennis, R.H. Problems in testing informal logic critical thinking reasoning ability. *Informal Logic, 6*, 3-9, 1984.

[13]Stewart, B.L. *Testing for critical thinking: A review of the resources*. (Rational Thinking Reports Number 2). Urbana, Illinois: The Illinois Rational Thinking Project, 1979.

Regarding observation appraisal Stewart discovered one test, *Recognizing Reliable Observations*, developed by the Instructional Objectives Exchange (IOX).[14] This test is based upon a set of principles of observation appraisal developed by Robert Ennis. We have modified this same set of principles for the observation test developed in this study. However, for the IOX test no validity or reliability information, other than a content validity judgement by expert opinion, is presented in support. Also Stewart points out,[15] and my examination confirms, that many of the reported observation statements on the test are not observation statements at all, but statements of inferences.[16] Another problem is that the test contains only ten multiple-choice questions meaning that the principles for judging observation statements, which far outnumber ten, are poorly covered, and that the test is likely to have poor reliability.

Others have explored the measurement of critical thinking ability, but their work has not received wide distribution nor are the tests which were developed readily available. In this regard work by D.P. Wright[17] and by the members of the Wisconsin Research and Development Center

---

[14] Instructional Objectives Exchange. *Recognizing Reliable Observations*. Los Angeles: Instructional Objectives Exchange, 1971.

[15] Ibid., p. 104.

[16] I have discussed the distinction between observation and inference in the following: A concept of observation statements. In D.R. DeNicola (Ed.) *Philosophy of Education 1981*, Normal, Illinois: The Philosophy of Education Society, 1982; A speech act conception of observation statements. In S. Clarke and R. King, *Papers from the sixth annual meeting of the Atlantic Provinces Linguistic Association*, St. John's, Newfoundland: Memorial University of Newfoundland, 1982; Defining observational competence, *Science Education*, *68*, 1984, 129-142; Observation in science and science education, Paper given at a conference on Heuristics in Mathematics and Science Education, Institute for Logic and Cognitive Studies, University of Houston, July, 1984.

[17] Wright, D.P. Instruction in critical thinking: A three part investigation. Paper presented at the annual meeting of the American Educational Research Association, New York, 1977, ERIC Document Reproduction Service No. ED 138 518.

BEST COPY

for Cognitive Learning[18] are important to mention because both employed principles of sound thinking in constructing the theoretical foundation for their work. This is the approach we have adopted also, since explicitly stating one's principles for good thinking guides test construction, curriculum development, and theorizing. In my mind, however, both pieces of work have serious flaws which the current study avoids.

Wright's work runs the risk of encouraging students to believe that the principles of thinking are exceptionless formulas for getting answers, and that judgements can be made when little of the relevant information is at hand. The work does this by requiring students to make judgements in situations in which no context is supplied. For example, one of Wright's questions is designed to assess whether or not students act in accord with the following conflict of interest principle: you get more reliable information from a person who has nothing to gain from his or her statement being correct than from a person who does. The students are to judge that Mrs. Truman's statement in the following example is reliable because she has no conflict of interest.

> Mrs. Truman teaches fifth grade. One day two fourth grade boys argued about who owned a ball. Mrs. Truman listened to Ted and Don and said, "I think it's Don's."

Given the small bit of information provided in the description of the situation a careful student would probably not conclude that Mrs. Truman's statement is reliable, and justifiably so. The most defensible response to this question would be to withhold judgement because not enough is known about Mrs. Truman's motives for ruling in Don's favour. We have attempted to avoid this problem in the test described in this study by setting the questions in the contexts of stories which provide the information required for making informed choices.

---

[18] Allen, R.R., Feezel, J.D., and Kauffeld, F.S. A taxonomy of concepts and critical abilities related to the evaluation of verbal arguments. Occasional paper no. 9. Madison, Wisconsin: Wisconsin Research and Development Center for Cognitive Learning, The University of Wisconsin, 1967. Eric Document Reproduction Service No. ED 016 658.

There are two major problems with the work done by the members of the Wisconsin Research and Development Center for Cognitive Learning: (i) the set of principles of sound thinking, consisting of only eight, is not comprehensive and (ii) some of the principles are stated as necessary conditions for reporting observations well, although they are not necessary conditions. For example, one of the principles says that in order to report observations well a person *must* be relatively free of bias or concern for personal benefit.[19] This is not so, however, since other factors such as the person's concern with reporting honestly might counteract the tendency to bias the report for personal gain. The principle would have been better stated had it said that a person's being likely to personally gain from his or her statement being correct *tends to* make that person's statement unreliable.

2.2. Critical Thinking Conceptualization

Much of the discussion of the previous section indicates that sound critical thinking test construction and test response interpretation depend upon having a comprehensive and sound analysis of the nature of good critical thinking. If one's analysis is not comprehensive, one obtains only a partial picture of people's critical thinking competence. If one's analysis is not sound, there is the risk of falsely categorizing those who are not good critical thinkers as being so, and those who are as not being so. In this study we have relied extensively upon Robert Ennis' analysis of critical thinking because in my estimation it is both sound and reasonably comprehensive, but also because it is in terms of principles of good thinking.

The Ennis approach of defining critical thinking in terms of guiding principles of thought has some distinct advantages. Consider the following

---

[19] Ibid., p. 17.

BEST COPY

principle as an example: "An observation statement tends to be believable to the extent that the observer has no conflict of interest." This principle, or more precisely the whole set of such principles of which this is just one,[20] has at least three practical uses. First, it provides as precisely as the subject matter will allow guidance for evaluating reports of observations. It says that the believability of such reports is diminished to the extent that the observers are in a conflict of interest. Note that this does not mean that being in a conflict of interest makes it impossible for an observer to make a credible report, as implied by the principle supplied by the Wisconsin Research and Development Center for Cognitive Learning. This would be too strict. It also does not provide any clear formula to indicate by how much conflict of interest reduces believability. Again, this is the only appropriate stance, since all such judgements of degree must be made in the context of some situation, taking into account all the information that is available. The principle says that *all other things being equal* the person in a conflict of interest is less credible than one who is not, and to this extent its guidance to evaluation is clear.

The second practical use of this principle, and all the others which go with it, is that they provide a core for a curriculum designed to improve students' thinking in this area. A presentation of the principles, a discussion of their meaning and intended mode of use, and provision for practice in applying the principles to the evaluation of statements of observation is the outline for a course on this aspect of critical thinking.

Thirdly, the set of principles provides the framework for designing a test of students' ability in this area, because it provides the justification for choosing correct answers on the test and a standard against which to judge the completeness of the test. The principles can be used for justifying test

---

[20]The whole set of principles and a discussion of them will appear in the following chapter.

answers because in a test one can effect the sort of control over the situation implied by the "All other things being equal" expression used above. That is, in a test one is able to contrive a situation in which all factors save one, say the observers' conflict of interest, are controlled. If one thinks of test items as experimental set-ups, and the giving of an item as running a trial of an experiment, as some do,[21] then all of this makes a good deal of sense.

The above considerations motivated me to base the test construction upon Ennis' conception of good observation appraisal. However, an examination of his set of principles in light of work in the philosophy of science and experimental studies of eyewitness testimony suggested that Ennis' set of principles needed some modification. This reanalysis was the first priority, before a draft test could be prepared. The resulting set of principles of appraisal is described in Chapter 3. In addition to these considerations, the nature of the study was influenced considerably by an attempt to use a long-touted but seldom employed method of test validation which relied on studying the mental processes of people while they responded to questions on the test. The motivation for adopting this approach, and the exact form it took on will be described in Chapter 4.

---

[21]Tomko. T.N. *The logic of criterion-referenced testing.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 1981.

Chapter 3

Principles for Appraising Observations

Public reports of observations are often subject to appraisal. Scientists must submit their observations to the examination of other scientists, eyewitnesses must submit their testimony to the evaluation of jurors. Assessing such reports of observation is a complex task, and there can be no exact formulas for how to do it properly. However, the greater one's knowledge of the factors which affect the accuracy of observation reports the greater one's ability to make sound assessments. When there are factors which influence reporting accuracy in a systematic way, then there is the possibility of codifying these effects and producing a set of principles of assessment based on their systematic occurrence. Table 3-1 contains the most comprehensive set of such principles that we have been able to produce thus far.

There are several things to notice about the set of principles. First, there are four categories. Principle I compares the believability of reports of observations to the believability of inferences based upon them. Principles II, III, and IV relate believability to characteristics of the observer, observation conditions, and the observation statement itself. Each of the latter three principles consists of several subprinciples, each addressed to a particular factor.

The principles must be interpreted cautiously. Specifically, it is not proper to treat them severally as either necessary or sufficient conditions

Table 3-1: Principles for Appraising Observations

I. Observation statements tend to be more believable than inferences based upon them.

II. An observation statement tends to be believable to the extent that the **observer**:

1. is functioning at a moderate level of emotional arousal;
2. is alert to the situation and gives his or her statement careful consideration;
3. has no conflict of interest;
4. is skilled at observing the sort of thing observed;
5. has a theoretical understanding of the thing observed;
6. has senses that function normally;
7. has a reputation for being honest and correct;
8. uses as precise a technique as is appropriate;
9. is skilled in the technique being used;
10. has no preconceived notions about the way the observation will turn out;
11. was not exposed, after the event, to further information relevant to describing it;
    (If the observer was exposed to such information, the statement is believable to the extent that the exposure took place close to the time of the event described.)
12. is mature.

III. An observation statement tends to be believable to the extent that the **observation conditions**:

1. provide a satisfactory medium of observation;
2. provide sufficient time for observation;
3. provide more than one opportunity to observe;
4. provide adequate instrumentation, if instrumentation is used.
   (If instrumentation is used in gaining access, then the statement tends to be believable to the extent that the instrumentation:
   a. has suitable precision;
   b. has a suitable range of application;
   c. is of good quality;
   d. works in a way that is well understood;
   e. is in good working condition.)

IV. An observation statement tends to be believable to the extent that the **observation statement**:

1. commits the speaker to holding a small number of things to be true;
2. is corroborated;
3. is no more precise than can be justified by the observation technique being used;
4. is made close to the time of observing;
5. is made by the person who did the observing;
6. is strongly believed to be corroboratable by the person making it;
7. does not conflict with other statements for which good reasons can be given;
8. is made in the same environment as the one in which the observation was made;
9. is not about an emotionally-loaded event;
10. is the first report of the event provided by the speaker;
11. is not given in response to a leading question;
12. does not report a recollection of something previously forgotten;
13. reports on salient features of an event;
    (Features of an event are salient to the extent that they are extraordinary, colourful, novel, unusual, and interesting, and not salient to the extent that they are routine, commonplace and insignificant.)
14. is based upon a reliable record, if it is based upon a record.
    (If an observation statement is based upon a record, then the statement tends to be believable to the extent that the record:
    a. was made close to the time of observing;
    b. was made by the person who did the observing;
    c. comes from a source having a good reputation for making correct records.)

for observation statements to be believable. That is, one or another of the conditions may not be satisified on a particular occasion but still a person may observe well. In addition, although one or another of the conditions may be satisified on a particular occasion, a person may still observe poorly. Thus, the expressions "tends to be believable to the extent that" and "tend to be more believable" used in stating the principles are meant to be taken seriously. They are qualifiers which indicate the limitations of the principles and stress that they are not exceptionless formulas. They must be applied judiciously, taking into account the characteristics of the situation at hand and relevant background knowledge, including experience in related matters. Application of the principles to actual cases is not a trivial matter, and requires an ability over and above knowledge and comprehension of the principles themselves. For example, in actual situations the principles often compete, pushing one's judgement of believability in different directions. In the case of two observers with different levels of access to the phenomenon, for instance, the one with the better access might also be the one with the lesser skill. Thus, an appraiser would have to weigh the different factors, taking into account the amount of skill needed for the phenomenon being observed and the comparative degrees of access which the observers have to the phenomenon. One might, depending on the situation, decide that good observational access is more important than high level of skill and conclude in favour of the observer with the better access. On the other hand, in a different situation the need for observational skill may far outweigh the need for good access, and an appraiser's judgement should change accordingly.

3.1. Justifying the Principles

Many of these principles are readily acknowledged by experienced, thoughtful people, and in that sense may seem trivial. However, research on eyewitness testimony appraisal[22] indicates that, in general, typical adults appraise observation reports using unsound principles. For example, jurors tend to evaluate eyewitness testimony according to such factors as the agreeableness of witnesses on the stand, their dress and physical bearing, their having cultured speech, and their expressing confidence in the truth of what they say. In addition, my own research on high school students[23] has revealed that they make consistent errors of reasoning, such as accepting things at face value when this is not warranted, relying unconditionally on the words of experts, and placing too much trust in the confidence which speakers display and in the definite-sounding manner of their statements. So while the principles may seem trivial to some, there are many people, including adults, who have very little knowledge of them.

The above argument points to the need for instruction in *some* principles of sound appraisal, but there remains the question of whether or not the particular principles in Table 3-1 are valid. Nelson Goodman[24] once argued that principles of inference, of which the principles in Table 3-1 are an instance, and particular inferences which we make are mutually justified by being brought into agreement with one another. That is, there

---

[22]Lindsay, R.C.L., Wells, G.L. and Rumpel, C.M. Can people detect eyewitness-identification accuracy within and across situations? *Journal of Applied Psychology, 66*, 1981, 79-89; Loftus, E.F. *Eyewitness testimony*. Cambridge, Mass.: Harvard University Press, 1979; Wells, G.L., Ferguson, T.J. and Lindsay, R.C.L. The tractability of eyewitness confidence and its implications for triers of fact. *Journal of Applied Psychology, 66*, 1981, 688-696; Yarmey, A.D. *The psychology of eyewitness testimony*. New York: The Free Press, 1979.

[23]Norris, S.P. and King, R. Observational ability: Determining and extending its presence. *Informal Logic*, forthcoming.

[24]Goodman, N. *Fact, fiction, and forecast*. Second Edition. Indianapolis: Bobbs-Merrill, 1965.

is an interplay between the two, each having the possibility to influence the other. He claimed: "A rule is amended if it yields an inference we are unwilling to accept; an inference is rejected if it violates a rule we are unwilling to amend".[25] The "we" in Goodman's directive is of course bothersome. If the "we" refers to the adult population at large, then the studies of eyewitness testimony mentioned above indicate that many patently false inference rules and inferences would have to be sanctioned. In addition, Stephen Stich and Richard Nisbett[26] have shown that this conclusion pertains to more than inferences about eyewitness testimony. Large numbers of the adult population make many sorts of unjustified inferences. These include improper assignments of probabilities to events (such as believing that the chances of flipping a head are increased after many tails are flipped successively), failing to account for statistical regression, and improper conclusions of causation from correlational data. Interpreting "we" broadly makes Goodman's rule dangerous, and thus for our purposes we modified the rule.

Stich and Nisbett suggested that Goodman's rule could be made usable by replacing the "we" with "the socially, consensually, designated authorities".[27] Part of the justification for our principles is provided by using Goodman's rule modified in this way. Many of the principles in Table 3-1 are based upon judicial practice. Rules of inference and of the admissibility of evidence have evolved in that field through the mutual adaptation of rules and particular inferences. Principles such as those dealing with conflict of interest, observer expertise and reputation for veracity, degree of access to the phenomena observed, leading questions,

---

[25]Ibid., p. 64

[26]Stich, S.P. and Nisbett, R.E. Justification and the psychology of human reasoning. *Philosophy of Science*, 47, 1980, 188-202.

[27]Ibid., p. 201

and record making, are examples of the type meant. If the judiciary is taken to be an appropriate, socially recognized authority in the sense above, then the use of these principles by the courts lends them credibility. In constructing the original set of principles upon which the one here is based, Robert Ennis[28] took a similar stance of relying on the practice of the courts.

Judicial practice is not, however, the only source of support for the principles in Table 3-1. In addition, there has been considerable research into the factors which affect the accuracy of eyewitness testimony, and generalizations about these effects can serve as the basis for principles. For example, this research consistently shows that witnesses report less accurately on emotionally-loaded events, ones in which there is violence, say, than on emotionally-neutral events.[29] Given the weight of the evidence we think it justified to include Principle IV.9: An observation statement tends to be believable to the extent that it is not about an emotionally-loaded event. Other principles supported by research from this field include those about reporting on salient features of an event, and being exposed to information relevant to the description of an event after the event has occurred. The principles are IV.13 and II.11, respectively. In these aspects of observation assessment, psychological research provides a supplement to judical practice. Even more strongly, the research provides grounds for changing certain features of judicial practice, and inferential practice in general.[30] The role extends beyond a mere description of the types of errors of inference and appraisal which people make, into

---

[28]Ennis, R.H. A concept of critical thinking, *Harvard Educational Review, 32*, 1962, 81-111.

[29]Loftus, Ibid.

[30]For example, the Law Reform Commission of Canada has made use of this research. See, Brooks, N. Police guidelines: Pretrial eyewitness identification procedures. Ottawa: Law Reform Commission of Canada, 1983.

providing rather direct information on how people *ought to* reason, and into how practice ought to proceed. For example, some psychologists are now using the results of their research to try to convince the judicial system that certain traditional modes of witness interrogation is likely to be counterproductive. Principle IV.10 provides a case in point. It is now common practice that witnesses must give accounts of their testimony on many occasions, to police officers, to lawyers, to insurance agents, and to the courts. However, evidence from eyewitness testimony research shows that people's accounts of what they have witnessed become less accurate as they give more and more reports. The accuracy is reduced to a larger degree than can be accounted for merely by removal in time from the event. Thus, there is reason to try to reduce the number of accounts of their stories which witnesses are asked to relate.

A third source of support for the principles derives from our common-sense psychology. Of course, one must be careful not to place undue emphasis on common-sense ideas, and be prepared to change them in the face of more systematic evidence. However, many of the principles are very plausible in light of common-sense notions and derive some of their support that way. For example, the principles about observer alertness, skill, and theoretical understanding, as well as those about adequate time and conditions for observing, derive partial support in this way. In each case we plausibly think of there being an accuracy-reducing mechanism in operation. Thus, we plausibly believe that if a person is not attending to an event then that person is less believable, because our common sense tells us that without attention memory traces of events are not stored or are stored inaccurately.

It is best, then, to think of the principles as being supported by a network of information. The practice of the courts, evidence from research on eyewitness testimony, and common-sense notions of the psychology of human beings serve jointly as both their source and their support. The

case is not the final word. New research and new understandings will likely lead to modifications. However, for now the set of principles is the most comprehensive and most accurate one available.

Chapter 4
Test Validation

The theory of test validation is a theory of how one ought to justify certain uses to which tests are to be put. If the use is as a measure of some psychological trait or *construct*, such as a mental ability, then the relevant validation technique is construct validation. The theory of construct validation is a theory of how to justify tests as measures of psychological traits. I take the trouble to say all this, and to risk repeating the obvious for many readers, for two reasons. First, construct validation played a central role in this study because the attempt was to produce a justified measure of a mental ability, the ability to correctly appraise observations. Second, the techniques of construct validation used in this study emphasized some approaches which are typically deemphasized in test validation studies, and placed less stress on those typically viewed as most important. Since this is so, this chapter is included to show why the approach taken here was adopted.

4.1. The Basis of a Theory of Construct Validation

I have argued elsewhere[31] that any theory of construct validation rests fundamentally on a theory of psychological constructs, or to put it another way, on a theory of the role of theoretical terms in science. None of this should be surprising. Psychological constructs are theoretical terms

---

[31]Norris, S.P. The inconsistencies at the foundation of construct validation theory. In E.R. House (Ed.), *Philosophy of Evaluation*. New directions for program evaluation, no. 19. San Francisco: Jossey-Bass, 1983.

in the same sense as terms such as "electron", "black hole", "gene", and our understanding of what we mean by these terms has to be fundamental to our conception of how science proceeds and what science produces. Specifically, a theory of theoretical constructs will determine to a large extent views of the nature of causal explanation, of the conception of truth, of standards of adequacy for judging theories, and of scientific theories themselves.

The problem that I perceive with construct validation theory as it is portrayed in the educational and psychological testing literature is that it is motivated by two inconsistent views of the nature of theoretical constructs, resulting in inconsistent views of causal explanation, of truth, of standards for judging theories, and of scientific theories. These inconsistent views are explicitly expressed in the literature, but they are not acknowledged. The result is that in essence there are two views of construct validation to be found in the test theory literature, though they are presented as one. Without giving textual support for the existence of these two views in the construct validation literature[32] I will outline the two views, and indicate how the validation methodology employed in this study is a result of adopting one of the views rather than the other.

4.1.1. Theoretical Terms

Theoretical terms in science have generally been interpreted in one of two ways. One of these, finding its roots in the work of David Hume[33] and

---

[32]This support can be found in the article by me mentioned above and also in Norris, S.P. A pitfall in the construct validation of ability tests, Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 1981; and Tomko, T.N. The logic of critierion-referenced testing, Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 1981.

[33]Hume, D. *An inquiry concerning human understanding.* C.W. Hendel (Ed.). Indianapolis, Indiana: Bobbs-Merrill, 1955. Originally published, 1748.

the philosophy of logical positivism expounded by The Vienna Circle,[34] views theoretical terms instrumentally. That is, they are useful instruments, useful for categorizing the observable world and relating the subsequently-produced categories to one another. The intended implication is that theoretical terms and the constructs to which they refer have no tie to the external world, except as ideas in the minds of scientists, that is.

In contrast to this positivistic view of theoretical terms, there is a view which maintains that the significance of theoretical terms in science is fundamentally no different from that of the terms we use for referring to everyday things, such as "chair", "water", and "grass". Part of the significance of these everyday terms is that we take them to refer to real things in the world: chairs, water, and grass in the cases above. To interpret scientific terms in this way is to assume that there is something real to which they also refer, although those real things are usually not directly observable, and often not likely to ever be. So, for example, to interpret the theoretical term "electron" in this way is to maintain a position much like the following one expressed by Hilary Putnam:[35]

> The statement that there are electrons flowing through a wire may be as objectively true as the statement that there is a chair in this room or the statement that I have a headache. Electrons exist in every sense in which chairs (or sensations) exist; electron talk is no more derived talk *about* sensations or "observable things" than talk about sensations or chairs is derived talk about electrons.

Applying these views to psychological constructs, such as ability terms, we find one view which maintains that abilities are no more than

---

[34]The Vienna Circle, [*The scientific conception of the world*]. In M. Neurath and R.S. Cohen (Eds.), *Otto Neurath: Empiricism and sociology*. Dordrecht, Holland: D. Reidel, 1973. Originally published 1929.

[35]Putnam, H. Three kinds of scientific realism. *The Philosophical Quarterly, 32*, 1982, 195-200.

classes of behaviours, and another view which maintains that abilities are unobservable, yet real, properties of people which give rise to their observable behaviours. Adopting one of these views is logically, and as it turns out empirically, linked to adopting related views on the nature of causal explanation, of truth, of standards of adequacy for judging theories, and of theories themselves.

4.1.2. Causal Explanation

Associated with his idea that theoretical terms have no reference outside the minds of scientists, David Hume proposed a theory of causal explanation which has followers to this day. The view is that causal connections, as real connections between things in the physical world, do not exist. Instead, causal connections are merely connections in the imaginations of people, stimulated by their having seen classes of events regularly "conjoined". Thus, to say that two events are causally related is to make a statement about the conceptual framework of people, not about some connection in the world. Hume argues for this position quite unequivocally:

> The first time a man saw the communication of motion by impulse, as by the shock of two billiard balls, he could not pronounce that the one event was *connected*, but only that it was *conjoined* with the other. After he has observed several instances of this nature, he then pronounces them to be *connected*. What alteration has happened to give rise to this new idea of *connection*? Nothing but that now he *feels* these events to be *connected* in his imagination, and can readily foretell the existence of one from the appearance of the other. When we say, therefore, that one object is connected with another, we mean only that they have acquired a connection in our thought.[36]

In contrast, there are philosophers who maintain that causal connections refer to real connections between the things causally related.

---

[36] Ibid., p. 86.

Rom Harre[37] argues that it is the task of science to discover the often unobservable, but nevertheless real, objects and processes which comprise these causal links. Contrary to Hume's view, Harre believes that causal connections must be more than conceptualizations in the minds of scientists, because a logical construction in the mind of a scientist cannot *make things happen*. And, Harre believes, whatever else is the nature of causes, they actually exert influence, they make things happen.

When applied to theorizing about psychological events, such as people's performances on mental ability tests, these views produce two versions of psychological explanation. In one, explanations of performances on these tests is given in terms of those performances being related in a consistent manner (conjoined) with performances on other tests or in other sorts of situations. The attempt is to produce a *nomological network*[38] which displays at least many of the relationships which investigators find pertinent. According to the other view, the attempt is to give explanations in terms of the mental mechanisms and processes which give rise to the test performances. The aim is to show what made the person perform as he or she did.

Often, these different sorts of explanations when given in sketches are not readily distinguishable. For example, it is not apparent from the explanatory statement, "The students' performaces were caused by their level of critical thinking ability," which sort of explanation is being assumed. There is a need to probe deeper to discover the sorts of support that is offered for the explanations. If the support is solely in terms of conjunctions of performances, the explanation is Humean. If the

---

[37]Harre, R. *The principles of scientific thinking.* Chicago: University of Chicago Press, 1970.

[38]Cronbach, L.J. Test validation. In R.L. Thorndike (Ed.), *Educational measurement*, Second edition. Washington, D.C.: American Council on Education, 1971.

justification is in terms of evidence of the mental processes and mechanisms which gave rise to the performances the explanation is of the type which Harre espouses. More will be said about this in the following section.

4.1.3. Standards of Adequacy for Judging Theories

For those who hold a Humean view of causation, theories are tested by deriving from them predictions of observable events and checking to see whether these predictions are upheld. To the extent that the predictions hold, the theories are supported; to the extent that they do not hold, the theories loose support. This view of scientific progress is called *hypothetico-deductivism* because it maintains that theories are first offered as *hypotheses* and then tested by *deducing* from them observable predictions.

While many who do not adhere to a Humean view of causation see a role for successful prediction in testing the adequacy of scientific theories, they realize that there must be other criteria of success. This is so since there are many theories which do not yield testable predictions. The theories of evolution and of plate tectonics, and archeological theories which attempt to explain unearthed discoveries, all receive support primarily (if not solely) from their ability to *explain* previously puzzling phenomena. A person who sees scientific adequacy in this light, draws a distinction between the conditions for successful prediction and those for successful explanation. Those who follow Hume's ideas typically conflate these conditions.[39]

Hypothetico-deductivism in one form or another has received

---

[39]Hempel, C.G. *Aspects of scientific explanation*. New York: Free Press, 1965; Keat, R. and Urry, J. *Social theory as science*. London: Routledge & Kegan Paul, 1975.

widespread endorsement in the test validation literature.[40] In the *Standards for Educational and Psychological Tests*[41] the following statement is made:

> In obtaining the information needed to establish construct validity, the investigator begins by formulating hypotheses about the characteristics of those who have high scores on the test in contrast to those who have low scores . . . Such hypotheses or theoretical formulations lead to certain predictions about how people at different score levels on the test will behave on certain other tests or in certain defined situations. (p. 30)

The *Standards* then indicate that confirmed predictions reflect favourably on the judgement of a test's validity and disconfirmed ones reflect unfavourably. This is hypothetico-deductivism in its explicit form.

4.2. Assumptions of the Validation Methodology Used in this Study

Hypothetico-deductivism was not the validation methodology employed in this study. When hypothetico-deductivism is appropriate at all (there are those who argue it is never appropriate)[42] it is so in highly advanced areas of science, when it is at least plausible that predictions can be deductively derived from theories. In the case of critical thinking ability, or the ability to appraise observations in particular, there is no available theory from which predictions can be derived deductively. This

---

[40]Campbell, D.T. and Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 1959, 81-105; Cronbach, L.J. and Meehl, P.E. Construct validity in psychological tests. *Psychological Bulletin, 52*, 1955, 281-302; Guion, R.M. On trinitarian doctrines of validity. *Professional Psychology, 11*, 1980, 385-398; Messick, S. The standard problem. *The American Psychologist, 30*, 1975, 955-966.

[41]American Psychological Association. *Standards for educational and psychological tests*. Washington, D.C.: American Psychological Association, 1974.

[42]Glymour, C. Hypothetico-deductivism is hopeless. *Philosophy of Science, 47*, 1980, 322-325.

is possible only when the relevant variables to be controlled are known. In the field of critical thinking ability it is not known, nor is there a theory of, how performances on critical thinking tests ought to relate to other performances. For example, it is not known how critical thinking ability, or any of its specific aspects, correlate with abilities in various school subjects or with performances outside the context of schools. Similarly, it is not known how the aspects of critical thinking ability relate, or ought to relate, to each other. Thus, though there are principles for appraising observation reports which concern three areas (the observer, the observation conditions, and the observation statement itself) it is not known, nor is there much reason to believe at this time, that knowledge of principles in one of these areas is related in a particular manner to knowledge of principles in the other areas.

The fundamental belief guiding the validation methodology of this study is that the essence of providing an explanation of some phenomena, say performances on a test, is providing a model of a mechanism (broadly construed) through which the phenomena were produced.[43] The essence of validating that explanation is attempting to find ways to discover whether that mechanism is indeed operative. For example, Robert Sternberg[44] has a model of a mechanism through which people solve analogies of the type found on the *Miller Analogies Test*. The mechanism is comprised of six components or steps. When one understands each of these components one can see that *if* this mechanism *were* operative in a person, and if there were no countervailing mechanisms operating at the same time, then that person could solve analogy problems of the type in question. Sternberg's research is largely devoted to attempting to discover whether this

---

[43]Bhaskar, R. *A realist theory of science.* Sussex: Harvester Press, 1978; Harre, R. *The principles of scientific thinking.* Chicago: University of Chicago Press, 1970.

[44]Sternberg, R.J. *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities.* Hillsdale, New Jersey: Erlbaum, 1977.

mechanism is indeed operative in people who do well on the analogy items, and with showing that other proposed models are less adequate because they fail, for one reason or another, to provide a mechanism which, if operative, would explain performance on the analogy items.

For the type of problems which examinees are presented on the *Test on Appraising Observations* I have proposed a mechanism which can lead to good performance. Each question on the test fits into an ongoing story. One of the stories is about a traffic accident and each question in this sequence presents two statements about the accident made by people who were involved or who were witnesses. The examinees are to pick which, if either, of the two statements they have more reason to believe at the time the statements are made. According to my theory of critical thinking, a critical thinker answering questions of this sort would approach them somewhat this way:

> The person would first of all look for relevant differences, either between the speakers (or if the same speaker is involved between that speaker's two occasions of speaking), between the conditions under which the observations were made, or between the types of statements that were made. If differences were found, their relevance would be judged according to whether or not there is some general tendency for such a difference to matter. For example, if one of the observers is a person who was involved in the accident and the other is a witness, then the examinee would note that the statement of the first is possibly influenced by a conflict of interest, and that people so influenced tend to be less believable than those who are not (all other relevant factors being equal). The examinee would then judge whether this general tendency concerning people in a conflict of interest applies in the particular situation, and if it does would reach the appropriate conclusion.

The above description outlines a mechanism through which a particular mental ability, the ability to appraise observations, can operate. If indeed those who do well on the *Test on Appraising Observations* do so because their performances are governed by a mental process such as this,

and those who do poorly do so because they use some improper approach, then the test would be judged a valid test of the ability to appraise observations. It is possible of course that other legitimate approaches might be used which result in good performance. Such a result could mean different things, depending on the particular mental processes involved. The researcher would have to examine the processes and determine whether or not they are of a sort that might be due to critical thinking ability. If they are, then despite the fact that they differ from original preconceptions, the test should be judged valid. If however they are not (they might be judged as guessing processes, or as processes relying on unintended cues in the items), then the test should be judged invalid.

All of this suggests that in order to judge the validity of the test there has to be some attempt to access the mental processes which determine people's performances on the test. The test is then judged valid to the extent that suitable mental processes lead to good performance and unsuitable ones lead to poor performance. The employment of a methodology for probing examinees' mental processes was at the heart of this study, and will be described in the following chapter.

Chapter 5

The Evolution of the Test on Appraising Observations

The *Test on Appraising Observations* in its present form (see Appendix A) is the result of an evolution through many versions. Each transition to a subsequent version was based on various considerations, some involving data which had been collected, others involving the best intuitions which we could muster. In this chapter, I describe in detail the basis of some of these transitions in order to illustrate more clearly the test development methodology that was used. Since what we were attempting was to provide at least a partial alternative to accepted test construction methodologies, it is important to give detailed descriptions.

5.1. Decisions Concerning Audience and Style

5.1.1. Audience

Before beginning any test construction one must have in mind (at least vaguely) the need which the test is going to fill. As mentioned previously, I have perceived a lack of critical thinking tests of particular aspects of critical thinking ability. So, the first need to be filled by this test was for a critical thinking test of appraising observations.

Beyond this decision, there are other matters related to whether the test is to serve classroom testing or research purposes (or both), or some other purpose. I decided to try to first fill the need of the classroom

teacher who often desires an easy to administer and score instrument, which can give rough indicators of students' attainment. Although detailed knowledge of students' capabilities is desirable, it is not possible in current classroom settings to have a running account of what students are learning at the detailed level. Teachers need frequent feedback at the general level of the success of their instruction, and this is best provided with short tests (one class period or less) which require only a few minutes to score per examinee.

Finally, one must decide on grade level. If one keeps the reading level of the test low, then one can produce a test which can be useful for a broad range of senior levels. For example, most tests which do not require the use of technical jargon can be written at the reading level of the average twelve or thirteen year old. So designed, the test can then usually be used at the junior and senior high school levels, and often at the university level as well. We decided to produce a test which was suitable for at least the senior high school grades, since this is probably the most likely place that techniques of observation appraisal would be taught. In so doing, the *Test on Appraising Observations* is probably usable in both lower and higher grades, although up until now it has not been tried there.

5.1.2. Style

Most concisely, the test can be described as a multiple-choice test whose items are set in the context of stories, requiring one class period to complete, and designed primarily for the high school grades. However, this description is somewhat misleading, especially regarding the categorization as a multiple-choice test. "Multiple-choice" seems to be the best description, but we have taken steps to avoid some of the most serious pitfalls which tests of this type have faced in the past. So for those who tend to have an immediate adverse reaction to anything which might be called a "multiple-choice test", I urge you to consider how this test is different.

The test consists of two parts with items set in the contexts of stories. Part A is the story of a traffic accident, and Part B involves the exploration of a river. The purpose of the stories is twofold. The primary one concerns the subject matter and the principles of observation appraisal which have been described. These principles are empirical generalizations, as was explained already. However, because they express *tendencies* and not *universals*, they admit of exceptions. In order to decide whether it is appropriate to apply one or another of the principles it is necessary to know the context in which it is to be used. This knowledge is necessary because extenuating circumstances can exist which can make application of a principle inappropriate. The contexts provided by the story lines in the test provide the required knowledge. Granted, they do not supply complete knowledge, which is tolerable since in no real situation is there ever complete knowledge. We must act on the best knowledge we can acquire. Providing some knowledge of context, however, avoids passing on the impression that the principles of observation appraisal can be mechanically applied in a knowledge vacuum.

The other reason for using stories is that they help to maintain interest in the test by keeping the examinees engaged in an evolving episode, in which one does not know what to expect next. We have found that this approach appears to have worked, because in informal conversations most students claim they did not find the test boring, although many said they found it puzzling.

The test contains 50 items (28 in Part A).[45] Although we have always given the test as a power test, forty-five minutes of actual working time is sufficient for most senior high school students to finish. Each item presents the examinee with two statements in bold type. They are to choose which, if either, of the statements in bold type they have **more**

[45]See Appendix A.

reason to believe at the time the statements are made. There are thus three alternatives for each item: (1) the first statement is more believable; (2) the second statement is more believable; and (3) neither is more believable, they are equally believable. In the sense that there are three choices for each item, the test is a multiple-choice test. However, the choices for each question are the same, which is not usual for multiple-choice tests. This has both advantages and disadvantages. One disadvantage is that in having only three distractors per item, compared to the more usual four, there is an increased chance that random guessing will lead to correct responses, thereby reducing the validity of the test. However, we have found that guessing on this test is in fact a rare phenomenon, and we can turn to our interviews of examinees to support this. Answers were virtually always chosen for reasons among the examinees we interviewed, and performances for those not interviewed were virtually identical to the performances of those interviewed. Thus, there is little reason to think that lack of guessing among interviewed subjects was a phenomenon caused by the interview situation.

Finally, though the number of distractors is usually greater than three for multiple-choice tests, when a test constructor is forced to find alternative distractors "to meet the required quota" there is a risk of having distractors of widely different attractiveness.[46] The result is that the *net* number of distractors is smaller than the *gross* number, because examinees with only minimal knowledge of the subject matter can often eliminate at least one distractor because of its obvious implausibility. The three distractors used in the *Test on Appraising Observations* are the logically obvious ones, given the problems posed in the test, and no evidence was found of their being eliminated as implausible alternatives.

---

[46]Lord, F.M. and Novick, M.R. *Statistical theories of mental test scores.* Reading, Massachusetts: Addison-Wesley, 1968, p. 309.

The final question of style concerns the nature of the problem as posed, namely, to judge for each question which, if either, statement in bold type is *more* believable at the time the statements are made. The decision to ask which is *more* believable was a deliberate attempt to avoid asking examinees to judge degrees of endorsement, such as would have been required had the distractors been, "strongly believe the first, strongly believe the second, weakly believe the first, weakly believe the second, believe each equally," or some other similar set. The motivation for doing this was the same that motivated Ennis in constructing the *Cornell Critical Thinking Tests*. The problem is that "people with different levels of sophistication justifiably give different levels of endorsement to a conclusion."[47] Asking, however, for only the direction in which the evidence points is a means of obtaining higher agreement among the best critical thinkers, and thus of more substantiated keyed answers.

5.2. Interaction of Validation Methodology and Stages of Development

A test construction is an exercise in *design*. Since this is so the designer's purposes and intentions play a fundamental role in the development and validation procedure. Although the designer need not be able to envisage all the uses to which an instrument can or will be put, he or she must have some purpose in mind, and all unforeseen uses of the test will derive at least part of their justification from the quality which was originally built in. In this respect I take exception to a remark by L. J. Cronbach which has received considerable endorsement in the testing field. He said:

> The phrase *validation of a test* is a source of much misunderstanding. One validates, not a test, but an *interpretation of data arising from a specified procedure*. A

---

[47] Ennis, R.H. Problems in testing informal logic critical thinking reasoning ability. *Informal Logic*, 6(1), 1984, 3-9.

> single instrument is used in many different ways -- Smith's reading test may be used to screen applicants for professional training, to plan remedial instruction in reading, to measure the effectiveness of an instructional program, etc. Since each application is based on a different interpretation, the evidence that justifies one application may have little relevance to the next.[48]

I cannot imagine how all of this can be true. It is true as I have already said that any test can be used for different purposes, some not even foreseen when the test was being designed. However, are all of these uses based on different interpretations, and is the evidence that justifies one interpretation irrelevant to the others? This is highly questionable. Consider the following example of a measuring device from the physical sciences. A voltmeter is designed to measure differences in electrical potential and to give valid readings (within specified limits of error) when used in a specified way. Specifically, the voltmeter must be used with the type of current (alternating or direct) for which it was designed; it must be connected in parallel with the electrical resistance across which the potential difference is being determined; the electrical resistance of the wires connecting the voltmeter to points in the circuit must be insignificant to the electrical resistance of the voltmeter itself; there must be no strong magnetic fields in the vicinity of the meter; and an unspecifiably large set of other conditions must be satisfied. Used in this way the voltmeter will measure voltage. However, just like Smith's reading test, the voltmeter may be used in many different ways. For example, it may be used to measure electrical current, or it may be used as an indicator of electrical circuit integrity. The voltmeter can be used to measure electrical current if the resistance across which it is connected is of known value. In such an instance, the known value of the resistance and the potential difference reading on the voltmeter can be used to calculate electrical current using

---

[48]Cronbach, L.J. Test validation. In R.L. Thorndike (Ed.), *Educational measurement* (Second edition). Washington, D.C.: American Council on Education, 1971, p. 447.

Ohm's Law, which is a functional relationship among potential difference, resistance, and current. The voltmeter can be used to indicate circuit integrity, if the only way in which a potential difference reading is possible in the circumstances is that there is a complete circuit. In this latter case, the voltmeter is not used to measure anything, because *any* reading indicates circuit integrity.

The question is whether each of these applications is based on a different interpretation, and whether the evidence that justifies one interpretation is relevant to the evidence that justifies the others. The applications are, it seems to me, based on different interpretations of the voltmeter reading. In the first, the reading is interpreted as a measure of potential difference; in the second, as a measure of electrical current; and in the third, as an indicator of circuit integrity. However, surely the evidence that supports the first application, the application for which the instrument was designed, is highly relevant to the two subsequent applications. In fact, the only reason that these subsequent applications are legitimate is that potential difference is functionally related to electrical current, and that to obtain a potential difference reading across a resistance there must be a complete electrical circuit present. That is, the subsequent applications and their justification are *directly parasitic* upon the intended application which guided the instrument design. There are, of course, uses to which the voltmeter can be put for which the intended interpretation is completely irrelevant. One might call to mind the high school physics student who, when asked to state how a barometer could be used to measure the height of a building, suggested as one answer that it be thrown off the roof and its time of descent determined!

Much the same can be said for test construction and validation. The test constructor, or to emphasize the intentional aspects of construction, the test designer, has an intended interpretation which is to be placed on the scores of the test. The aim is to design a test for which this

interpretation is legitimate. The initial validation of the test is an appraisal of whether this aim has been reached. If subsequently the test scores are to be interpreted differently from the originally intended interpretation, then the legitimacy of these interpretations must be determined. The original validation study will, however, be relevant in this regard, just as the validation of the voltmeter as a measure of voltage is relevant to its use as a measure of current or as an indicator of circuit integrity.

The view guiding this study is that validity investigations of ability tests involve three steps: (i) finding out whether examinees understand the tasks on the test in the intended way; (ii) given that the appropriate understanding is there, determining whether examinees use appropriate approaches to complete the required tasks; and (iii) if number of correct answers is to be used as the indicator of examinees' ability, determining whether the keyed answers are justified given appropriate strategies for arriving at answers. A word on each step is in order.

If examinees understand the tasks they are being asked to do in a way different from that intended by the test constructor, then there is a risk that the intended interpretation of test scores will not be legitimate. Thus, if examinees taking the *Test on Appraising Observations* interpret the instruction to determine which statements, if either, are more believable as an instruction to determine which statements are true, then this might affect the legitimacy of interpretations of their scores. The situation is not straightforward. One would have to examine the basis for choosing answers. This examination might indicate that any interpretation of scores in terms of ability to judge comparative believability of reports of observations is ill-founded. On the other hand, the examination might reveal a discrepancy only at the word usage level, and not at the level of underlying meaning.

Given that the examinees understand the tasks in the intended way, it is not necessary that all examinees use the same approach to solve the tasks. The test designer may have an approach in mind, but must be open to there being other legitimate approaches to completing the same tasks. The appearance of such alternate approaches does not of itself mean that the test is invalid. However, since they cannot always be anticipated in advance (there is no way to predict the approaches of ingenious examinees) such approaches must be examined for their legitimacy at the time they appear.

Given the intended understanding and the use of appropriate approaches on the part of examinees, it must be determined whether the keyed answer is appropriate. It is not unusual to find that examinees can follow approaches different from those anticipated by the test designer (yet appropriate all the same) to reach answers different from those keyed. The most usual result of this occurrence is the abandonment of the troublesome items.

The general principle of ability test validation underlying this study is that ability tests are valid to the extent that good thinking leads to good performance on the test and poor thinking leads to poor performance. The attempt is thus to explain performance in terms of thinking, and to do this there must be a description of the thinking processes which lead to performance. It is important to note that this explanation need not be applicable to all examinees for the test to be suitable. It is because it is always possible, given our current lack of ability to identify and control the relevant variables, for good thinking to result in poor performance and for poor thinking to lead to good performance that the expression "to the extent that" was used above. For a test to be suitabl, valid, there must be at least an overall tendency for good and poor thinking to be linked to good and poor performance respectively.

The approach to validating the *Test on Appraising Observations* was guided by the above general principle. Each stage of development was an attempt to bring the test closer to the ideal of having good thinking being the sole cause of good performance and poor thinking the sole cause of poor performance. Actually, the latter is more difficult for the test designer to influence, because poor performance can result from so many completely uncontrollable influences, poor attitude, lack of health, etc. The following section describes the stages of development of the test.

5.3. Transitions Between Test Versions

The current version of the *Test on Appraising Observations* is the result of the modification of several previous versions, some highly experimental which were used to chart rough directions, and some highly refined which required only cosmetic changes. I will describe each version, while concentrating on the most crucial points in the test's evolution.

5.3.1. Preliminary Versions

The earliest version of this test which was administered to a substantial number of subjects contained two parts and seventy items, fifty-nine in Part A.[49] There was an obvious imbalance in numbers of items in each part of the test which was subsequently altered. However, the test had taken on its basic form, with two story lines and three-alternative questions.

This test was given to 51 sophomores in a public high school in central Illinois. The average score on the test was 37 items correct, with a KR-21 reliability of 0.75. Both of these results were judged respectable for a trial version, as were comments from students in discussions held with

---

[49]Norris, S.P. Illinois Test on Assessing the Reliability of Observation Statements, Illinois Rational Thinking Test Series, Bureau of Educational Research, University of Illinois at Urbana-Champaign, July, 1979.

the classes afterwards that they found the test engaging. However, certain obvious problems emerged. The instructions in this preliminary version told examinees to choose which, if either, statement in each question they found more dependable. Many students were not sure what was being requested, and attempts to solve their uncertainty with synonyms such as "reliable" and "trustworthy" did not seem to help. In the subsequent version the instructions told students to choose which, if either, statement they had more reason to believe. This has proved satisfactory, and the instructions have remained the same in all versions since.

Another obvious problem was that the test was too long. Students said that the test required them to work very hard, and near the end they were too tired to concentrate well. Also, only one-half the students finished the test and only three-quarters finished the first part. This was unacceptable for a power test. Test length was reduced to fifty items in the current version.

There were also about a dozen items which had suspicious discrimination characteristics. These were items for which getting an item correct was negatively or negligibly related to overall performance on the test. Each of these items was examined and changes were made where the problems seemed to be occurring.

A modified version of the test was then given to 94 public school students in another town in central Illinois. This version had not been shortened, and the effect was similar, with only about one-half the students finishing. However, the KR-21 was substantially higher, 0.85, and many of the items that formerly had suspicious discrimination indicies looked much better. However, there was still an obvious need of improvement. The test length problem had to be corrected. In addition, the principles of observation were not at that time as widely based as they currently are. More work was required here, which led to a need for different items since new principles were added.

A final preliminary version was thus prepared which was based on an improved set of principles of appraisal (those in Table 3-1), which had a better balance in length between Parts A and B, and which was substantially shortened. The test contained 53 items in all with 25 in Part A. This version was administered to a small sample of 11 high school students from St. John's, Newfoundland, chosen for their availability during the summer months and given a small honorarium for their assistance. Each student was asked to take just one part of the test (6 took Part A) and was asked to think aloud while doing it. This procedure was beneficial in several ways. It indicated places where items or instructions were ambiguous, and suggested how subsequent interviews might be conducted. The first experimental version of the test, to be used to collect think aloud protocols from a large number of students, resulted from making modifications to this final preliminary version.

5.3.2. Experimental versions

The first experimental version of the test, called *Test on Assessing the Believability of Observation Statements, Version B* (see Appendix B), was a 50 item test of the same format as the previous version. Three questions were dropped from the previous version, and 15 others were altered or replaced to accomodate the deficiencies which had become apparent in the 11 interviews and to effect a more even distribution of items across the set of principles. We were thus ready to collect the desired interview data.

Our desire was to conduct the interviews in a fundamentally non-leading fashion. We wished to influence students' thinking as little as possible, realizing that just asking people to think aloud and placing them alone with a stranger might have effects in themselves. At the same time, it seemed that sometimes interrupting a student's narrative might be more beneficial than not, particularly when the interruption was merely to

clarify the ambiguous referent of a pronoun, or to point out obvious reading errors. In addition, although we did not wish to rush the examinees, to cut off reasoning by inadvertent signals, or to endorse or criticize particular reasoning attempts, we did wish to obtain as complete records of reasoning as were possible. To fulfill this aim it was often necessary to probe beyond the initial instruction to think aloud. This probing was done, however, only after examinees had chosen their answers to questions and had finished reporting on their thinking. Even in these follow-up stages, probing was as non-leading as possible, merely echoing examinees' already reported thoughts or asking them whether they could explain a little more about their choices of answers. The *Observation Test Interview Model, B* (Appendix C) indicates four stages of the interview process, ranging from the least leading to the most leading. The first stage merely informed the examinees of the general purpose of the interviews and that they would be asked some questions while taking a test.

Stage Two consisted in the first level of probe into examinees' thinking, and was non-leading. Examinees were asked only to tell all they could about what they were thinking as they were choosing their answers. This stage permitted interruptions only to probe for ambiguous references in examinees' reports, and for obvious reading errors. There was no provision for answering examinees' requests for additional information or for feedback on their progress. The interviewer was permitted to give only uninformative responses to such requests: "You can only go by what is written", or "You can decide only according to what is said and what you know."

The third stage of interview was more leading and the particular probes used depended both upon the answers chosen and upon the thinking reports. We had developed a model of an ideal answer for each question which involved first identifying the criterion or factor which made the difference in each case (one speaker was in a conflict of interest, one

speaker had more expertise, the observation conditions were better in one case than the other, etc.), then using this criterion to make a comparison between the two statements in the question on the basis of a general principle explaining the relevance of the criterion. If no identification of the criterion appeared in examinees' initial report, then the criterion was identified by the interviewer and the examinees were asked whether the criterion played any part in their thinking. If examinees mentioned the criterion but went no further in explaining the relevance of the criterion to their choice of answer, they were asked to tell (if they could) more about the difference the criterion made. For example, if in number 6 an examinee said that he or she chose the first statement (See Appendix B) because Ms. Vernon was a driver education instructor, then the student would be asked: "Could you tell me more about the difference Ms. Vernon's being a driver education instructor makes to your thinking?" Finally, if an examinee mentioned the proper criterion and also explained its relevance to the chosen answer but did not show how this relevance was derived from some general principle, the examinee was probed very indirectly so as to be given a chance to add more information. So, if a student mentioned that Ms. Vernon's being a driver education instructor and thus being more expert than Martine makes the difference, the examinee was probed: "So *Ms. Vernon's being a driver education instructor* makes the difference?" If an examinee merely responded "Yes", then he or she was asked to go on to the next question. If there was further explanation, this was recorded.

The interview data was then used to rate the quality of examinees' thinking. This information was coupled with the answers they had chosen to make judgements about whether or not particular items and the test as a whole had worked properly. Based on this feedback yet another version of the test was produced, Version C (See Appendix D), which was subjected to analysis through interview data collected in the same way as

that described above. When this stage had been reached the test seemed in satisfactory condition, except for a few minor cosmetic changes. These were made and the final version (See Appendix A) was administered to a large sample of students in order to acquire stable reliability estimates and sufficient data for producing some norms.

5.4. Summary

This chapter has provided an overview of the developmental stages of the *Test on Appraising Observations*. The test has evolved through many stages, initially more through intuitive judgements of what seemed sound, later through data collected on small numbers of students, and finally through rigourously collected and analyzed data on relatively large numbers of students. The aim was always the same: to produce a test for which good thinking generally led to good performance and poor thinking generally led to poor performance. The desire to have rather direct evidence on whether or not this aim was being met controlled more than anything else the methodology that was used. The systematic collection of think aloud protocols, their analysis, and the changing of items based on this analysis was a direct response to this aim. We believe that the *Test on Appraising Observations* meets this standard to a satisfactory degree. The degree to which it does is documented in the following chapter.

Chapter 6

Data Collection and Analysis: Final Results

In this chapter we describe in detail the data collection and analysis for the two experimental versions of the test, and show how the data was used in the transitions from Version B to Version C and from Version C to the final version. In addition, item and test statistics for the final version of the test are given.

6.1. Samples and Data Collection

Data on Version B was collected in the Fall of 1982 from two samples from seven senior high schools on the Avalon Peninsula in eastern Newfoundland. The communities varied from relatively isolated to relatively urban. The first sample was of 181 students in levels I and II (at the time Newfoundland did not have a three-year high school) and the second of 52 students from the same grade levels. The first sample was chosen on the basis of intact classes. Each class was administered Version B of the test and asked to try to finish the entire instrument. Eight classes in all, four in each of levels I and II , were given the test in this way. The second sample consisted of students randomly chosen from classes in the same schools which had not been administered Version B, and randomly assigned to one of us for testing. Each student in this sample was asked to take either Part A or Part B of the test, and to take the first half of the part assigned in the interview format and the second half in the normal testing format. Table 6-1 illustrates the sampling format.

Table 6-1: Distribution of Samples by Grade Level, School, and Testing Format

	Group Testing		Interview Testing	
School	Level I	Level II	Level I	Level II
1	X			X
2	X	X	X	X
3		X	X	
4		X	X	
5	X			X
6		X	X	
7	X			X

Data on Version C was collected in the Spring of 1983 from two samples of students from four schools in eastern Newfoundland. Two of the schools were in relatively urban settings while the others were in rural areas. The first sample which administered the entire test in intact classes and consisted of 171 students in levels I and II of the senior high school. The second group was interviewed individually on only one part of the test and consisted of 44 students.

Data on the final version of the test was collected in the late Spring of 1983 from four senior high schools in southern Ontario. Students were administered the test in intact classes and in addition to the *Test on Appraising Observations* (TAO) were administered either the *Cornell Critical Thinking Test, Level X* (CCTT) or the *Watson-Glaser Critical Thinking Appraisal, Form A* (W-G). The tests were administered through

the cooperation of Professor Philip Nagy who provided guidance to the schools involved. Each cooperating teacher received a copy of the instruction sheet in Appendix E. The testing was done through the schools' science departments in biology, chemistry, and physics classes. Consequently, the sample was biased towards students of average to above average achievement. Sampling was as shown in Table 6-2.

Table 6-2: Number of Students in Ontario Samples for TAO, CCTT, and W-G

School	TAO	W-G	CCTT
Bluevale	101	45	42
Forest Heights	287	129	153
Guelph	100	65	36
Waterloo	106	34	42

6.2. Basic Data and Derived Data

6.2.1. Basic Data

The basic data for this study came from two sources, students' choices of answers to questions on the tests and thinking protocols collected during interviewing. Answers to questions were recorded by students on standardized answer sheets for each of the tests. The interviewed students typically verbalized their answers in addition to recording them on their answer sheets. If the verbalized answer differed from the recorded answer (which was the case only seldomly), then no remark was made by the interviewer. We wished answer selection errors to be present in our data since in normal testing situations they cannot be eliminated. Thinking

protocols were tape recorded and in addition we made brief notes while students were thinking aloud. These notes were afterwards checked for accuracy against the tape recordings. Finally, from the answer sheets and the think aloud protocols a number of variables were derived which served as the basis for evaluating the quality of the test.

6.2.2. Derived Data

Eight variables were derived from the basic data collected in the study: performance scores, thinking scores, item/test biserial correlation coefficients, item difficulty levels, item-thinking/item-performance biserial correlation coefficients, item thinking/performance index scores, respondent thinking/performance index scores, and Kuder-Richardson-20 reliability indices.

Performance scores. Performance scores were given as number of items answered correctly according to the accepted key. No correction was made for guessing. For various analyses it was advantageous to derive performance scores for particular parts of the test, either Part A or B, or for the part on which students were interviewed and the part completed before or after being interviewed.

Thinking scores. The test was divided into four logical divisions according to the story line; Part A, items 1-15 and 16-28, and Part B, items 29-37 and 38-50. Students were interviewed on one of these sections, thus having to report their thinking on from 9 to 15 questions. Thinking scores were based upon an analysis and evaluation of these reports. For each item, thinking was rated according to the scale in Table 6-3. Table 6-4 shows how a student's thinking might be rated for item 3 of the test.For each item students could receive a thinking score of between 0 and 3, and depending on the part of the test being taken could receive a total thinking score for that part of between 27 and 45. The effective maximum

thinking score per item was, as it turned out, 2. Respondents rarely thought so well as to receive 3 for an item. Thus, total thinking scores realistically had maximums of 18 and 30 depending on the part of the test.

Table 6-3: Rating Scale for Thinking Scores

Rating	Basis of Evaluation
Rating = 1	the respondent cites the criterion by which correct appraisal of the underlined statements may be made, or the respondent uses the criterion in comparing the two underlined statements but does not explicitly cite the criterion
Rating = 2	the respondent cites the criterion and also uses it to compare the two underlined statements
Rating = 3	the respondent cites the criterion, uses it to compare the two underlined statements, and also generalizes from the particular situation to situations like it
Rating = 0	the respondent does none of the above or does not respond

Item/test biserial correlation coefficients. For each item, the biserial correlation between item performance (right or wrong) and total score on the test was calculated. This statistic was calculated only for those students who were in the non-interview groups and who thus completed all sections of the test.

Item difficulty levels. Difficulty levels as measured by the proportion

Table 6-4: Rating Thinking for Item 3

Rating	Basis of Evaluation
Rating = 1	The student points out that Mr. Wang was involved in the accident.
Rating = 2	The student points out that Mr. Wang was involved in the accident, and compares Mr. Wang's involvement with Ms. Vernon's being a bystander.
Rating = 3	The student points out that Mr. Wang was involved in the accident, compares this with Ms. Vernon's non-involvement, and shows how this is an instance of a more general phenomenon. For instance, the respondent might say that people involved in situations where they might be blamed tend to be less believable than those who cannot be blamed.

of students getting items correct were calculated for each item. Since this index was given as a proportion of respondents getting an item correct, the lower the index the harder the item was assumed to be for the group as a whole. Note the technical definition of difficulty in terms of overall group performance, however, since it may not coincide with everyone's concept of difficulty. For example, there is no necessity to conceive item difficulty such that greater difficulty leads to poorer performance. For example, one might imagine greater difficulty leading to greater motivation and thus to better or equal performance. Like the item/test biserial correlation coefficients this statistic was calculated only for those students who had completed the entire test.

Item-thinking/item-performance biserial correlation coefficients. The

biserial correlation between item thinking scores and item performance scores was calculated for each item across subjects. This produced an index for each item of the relationship between thinking well on that item and getting the item right. One problem with the coefficient is that in the extreme (and in cases close to the extreme) when all (or most) subjects receive 0 for thinking as well as 0 for performance on an item, the biserial correlation yields an index of 0. However, in our opinion an item with a majority of (0,0) scores for (thinking, performance) is working well. When people think poorly they get the item wrong. Furthermore, in certain cases the biserial correlation coefficient is greater than unity, presumably because certain assumptions are violated by the data, normality being one of the most important assumptions for this statistic. To help capture the relationship between item thinking and item performance the index described in the following section was devised to be used in conjuction with the biserial correlation.

Item-thinking/performance index scores. In order to try to offset somewhat the deficiencies of the item-thinking/item-performance biserial correlation, a thinking/performance index (T/P index) score was developed. First of all, combinations of thinking and performance scores were rated as in Table 6-5. Thinking scores from 0 to 2 only were chosen since there were so few scores of 3 for thinking. Any thinking score of 3 was thus converted to 2. There were thus six possible combinations of thinking scores and performance scores, (T,P) scores. Combinations (0,0) and (2,1) were judged to give the same degree of positive evidence for the quality of an item (it being assumed as discussed in the chapter on test validation that one sign of quality items is that good thinking leads to correct responses and poor thinking to wrong responses), and were assigned a rating of +2. Combinations (0,1) and (1,0) were judged to provide the same degree of negative evidence for the quality of an item, and were assigned a rating of -1. (2,0) was judged to provide a higher degree of

negative evidence than either of the previous two combinations and was assigned a rating of -2. Finally, the combination (1,1) was judged to provide an intermediate level of positive evidence for item quality and was rated +1.

The aim was to provide an index of the average degree of evidence from different sources on the quality of an item, not to provide an index of correlation as with the biserial coefficient. Thus, to arrive at the thinking/performance index score for an item, the scores for all subjects answering an item, as determined by Table 6-5, were averaged and the result divided by 2. The index thus ranges from -1 to +1, with -1 indicating the highest level of negative evidence against the quality of an item, +1 the highest level of positive evidence, and 0 that the positive and negative evidence was in balance.

Table 6-5: Weights of Evidence for Thinking/Performance Combinations

		Thinking Scores		
		0	1	2
Performance	0	+2	-1	-2
Scores	1	-1	+1	+2

Respondent thinking/performance index scores. Once thinking/performance evidence weightings per student and per item were calculated, there were several possibilities of derived scores. One was the item thinking/performance index score just discussed. Another is a thinking/performance index score for the respondent, calculated by averaging evidence weightings obtained by that respondent for all items on

the test. This index gives a rating of the degree of evidence for the quality of the test as a whole as obtained from that one respondent.

Kuder-Richardson-20 reliability indices. KR-20 reliability estimates, giving a lower bound on the reliability of the test, were calculated using as data the responses of those who completed all items on the test.

6.3. Analysis: Version B to Version C

The primary concern in all of our analyses was to determine the extent to which test performance could be explained by level of critical thinking, and adjustments were made in situations where there were systematic tendencies for some other factor or factors to explain performance. For this reason, considerable weight was placed on the thinking/performance index scores. Other indicators were considered but always in light of the thinking/performance relationship. In particular, the following questions were addressed and changes were made in light of the answers found:

1. What was the relationship between thinking and performance for each item?

2. What was the relationship between thinking and performance for the test as a whole?

3. How does performance compare on items measuring the same principle?

4. How is performance on each item related to overall test performance?

5. Was "test wiseness" a factor?

6. Did the test systematically mislead in any way?

7. Was reading difficulty a factor?

8. Were the instructions clear?

9. Were any of the test's characteristics systematically related to the ability of the students?

10. Did any of the factors, interviewing, interviewer, test section, grade level, or sex of examinee, affect thinking scores?

11. Did any of the factors, interviewing, interviewer, test section, grade level, or sex of examinee, affect performance scores?

The information contained in Table 6-6 as well as the student protocols served as the primary information for answering the above questions.

6.3.1. Thinking/Performance Relationships for Items

The relationship between thinking and performance for each item was indicated by both the item-thinking/item-performance biserial correlation coefficient and by the item thinking/performance index score. It is a matter of informed judgement, however, which magnitudes of these numbers should signal a warning. An obvious signal occurs if either of the numbers is negative. A negative biserial correlation would indicate that for that item poor thinking tends to lead to correct performance and good thinking to incorrect performance. A negative thinking/performance index score would indicate that there is more evidence against the quality of the item than there is in favour of it.

Beyond this, however, it is difficult to know what should be taken as a low T/P index score or a low biserial correlation. A lower bound of acceptability for the T/P index can be derived on the assumption that a T/P index score should at least be as high as the score which would obtain for a sample of students randomly guessing their answers and all thinking poorly. In this situation, one-third of the students would get the item correct and all students would receive zero for thinking. For a sample of N students, N/3 would have a (T,P) combination of (0,1) yielding an evidence weight of -1 from Table 6-5, and 2N/3 would have a (T,P) combination of

Table 6-6: Item Level Statistics, Version B

Item No.	Item/ Test biser.	Diff. Level	T/P biser.	T/P index	Item No.	Item/ Test biser.	Diff. Level	T/P biser.	T/P index
1	.131	.702	.622	.462	26	.538	.449	.527	.333
2	.185	.757	.936	.731	27	.292	.469	.344	.413
3	.354	.425	.787	.616	28	.365	.299	.902	.750
4	.301	.287	.689	.615	29	.067	.188	.0	.875
5	.186	.514	1.169	.885	30	.248	.335	.551	.423
6	.296	.547	1.063	.769	31	.349	.466	.460	.346
7	.182	.249	.206	.577	32	.358	.577	.907	.692
8	.437	.558	.568	.385	33	.222	.431	.838	.654
9	.150	.276	.867	.769	34	.462	.401	.846	.538
10	.417	.751	.623	.346	35	.428	.593	.574	.769
11	.074	.398	-0.727	-.231	36	.403	.494	.623	.461
12	.201	.354	.717	.538	37	.520	.655	.999	.731
13	.282	.238	.451	.458	38	.211	.456	1.296	.334
14	.374	.331	-0.162	.208	39	.350	.314	1.006	.833
15	.123	.215	.863	.625	40	.428	.605	.829	.625
16	.313	.348	.654	.536	41	.333	.333	-0.322	.250
17	.409	.628	.552	.715	42	.459	.491	.269	.375
18	.447	.556	1.338	.964	43	.433	.401	.556	.209
19	.411	.606	.927	.786	44	.264	.184	1.006	.875
20	.327	.539	.034	.250	45	.400	.384	.723	.667
21	.417	.472	.149	.250	46	.501	.685	1.240	.958
22	.299	.228	.430	.429	47	.476	.493	1.264	.833
23	.419	.263	.420	.429	48	.448	.636	1.192	.917
24	.290	.302	.0	.679	49	.475	.588	.848	.708
25	.488	.475	.959	.769	50	.329	.176	.886	.750

(0,0) with an evidence weighting of +2. Such an occurrence would give a T/P index score of .5. Note that this is a reasonably high score since from two-thirds of the sample the highest possible evidence weighting of +2 is obtained. Thus, choosing a T/P index of .5 as a lower bound of acceptability is quite conservative.

An examination of Table 6-6 shows that to one decimal place there were 16 items with T/P index scores less that .5 and one, item 11, had a negative score. Each of these items was treated as suspect and were examined in light of the other indicators of quality and in light of the protocols. In the end, all but two of the items, numbers 22 and 23, were altered. These items were close to the standard of acceptability in any case, and we thought that changes which were made to neighbouring items 20 and 21 would make 22 and 23 stronger items as well. The strategy was to try not changing them, since no obvious way of changing them was apparent, and to see how they behaved in the next trial.

The changes that were made to items ranged from single word changes, to the addition or deletion of material, to changes in the instructions of the test. As a result of the protocols and T/P index scores extensive changes were made to the introduction to the traffic accident story. In the Version B Introduction the names of all the characters and what they were doing at the time of the accident were provided. Since there were several characters in the story with different roles, we thought that providing the names in a single list would assist students in keeping them straight. However, there were unexpected problems caused by doing this. For the first six items many students referred to the introduction for evidence to support their choice of answers. While this is a legitimate thinking strategy, it contributed to uncontrolled influences on students' responses and thence to unjustified interpretations of thinking from performances. For example, in item 1 the keyed answer is that Martine's statement that there were three cars at the scene of the accident is more believable. Good critical thinking would lead to this response because Martine who was driving would tend to be more alert to the number of cars than Pierre who was reading a map and trying to figure out which way to go. However, eight students chose the correct answer by referring to the introduction and counting the number of cars mentioned there. One student reasoned as follows:

> [The first statement] is more believable because Martine was in the car, too and well, when I read back there [pointing to the introduction], well, there were three cars, so she would have to be right. Maybe Pierre, he said there was five cars. Maybe he was doing something else at the time of the accident.

Now as it happens the introduction does not say how many cars were at the intersection, but does mention three cars which were involved in the accident. Thus, a student who was not thinking critically would assume that the number of cars at the intersection equalled the number of cars in the accident, whereas the critical thinker would realize the fallacy of this reasoning. Thus, the non-critical thinker would be rewarded with getting the item correct through an unsound reasoning process.

The T/P index score for item 1 indicates that this item seemed to work satisfactorily despite the type of influence just described. However, other items fared less well. For example, for items 11 and 14 (T/P indexes of -.231 and .208 respectively) some students still referred to the introduction for information which led them to choose the correct response but not for the desired reasons. This contributed to the poor behaviour of these items and provided more motivation for altering the introduction.

It can be seen from the above examples that the computed relationships between thinking and performance were used in conjunction with the protocols of students' thinking. If the relationship between thinking and performance was low, a reason for this was often found in the protocols, which provided sound information on the types of corrective measures to take. This marks a substantial improvement over trying to make changes to items based on item/test biserial correlations and item difficulty levels without information on those factors which influenced students' responses.

6.3.2. Thinking/Performance Relationship for the Test as a Whole

The thinking/performance relationship for each item is obviously related to this relationship for the test as a whole. However, the latter statistic is useful in its own right. Table 6-7 gives the correlation between thinking scores and performance scores for the entire test, and for both parts of the test separately.

Table 6-7: Pearson Correlations Between Thinking Scores and Performance Scores

Test Section	Pearson's r	r^2	Significance
Whole Tes	0.68	0.46	>.001
Part A	0.59	0.35	>.001
Part B	0.77	0.59	>.001

The correlations obtained were judged satisfactory, at least at this stage in the test development. They were of the same order of magnitude as the KR-20 reliability which was computed to be 0.72. The relatively lower correlation obtained for Part A compared to Part B was explainable in terms of the considerable difficulty which was caused in Part A by examinees referring to the introduction for assistance. This was not a problem for Part B. Thus, no changes were made to the test based on this information, but a caution was registered to check for an imbalance between the two parts in subsequent versions.

6.3.3. Items Testing the Same Principle

As described previously, each item on the test was designed to test one of the principles in Table 3-1. While we were not overly concerned with internal consistency in the test as a whole, as measured by the KR-20, say, we judged that performances on at least those items measuring the same principle should correlate well. It is reasonable to argue that the ability to appraise observations is multi-dimensional and that the various aspects need not correlate well. However, this argument is not available for items testing the same principle. At the same time, it must be realized that even for items measuring the same principle, there was always a slightly different context. It is known that context affects reasoning substantially,[50] so we cannot expect the correlations to be perfect even were there no invalidating influences operating.

Instead of computing correlations between items testing the same principles we chose to examine the item difficulty levels and the item/test biserial correlations. For the above reasons we did not demand that these be nearly equal nor did we automatically alter items just because they were substantially different. Instead, when the figures differed substantially (by about 50%), we took that as a reason to treat the item suspiciously. Table 6-8 gives the item difficulties and the item/test biserial correlations for items designed to test the same principle.

As noted in the table there are a total of 6 principles for which either the difficulty levels or the biserial correlations of their items differ by 50% or more. This involves 14 items in total, of which 10 were altered for the subsequent version. The remaining 4 items were unchanged because we felt that changes made elsewhere would remove the problems with these items.

---

[50]Evans, J. St. B.T. *The psychology of deductive reasoning.* London: Routledge & Kegan Paul, 1982.

Table 6-8: Item Difficulties and Item/Test Biserial Correlations for Items Testing the Same Principles

Principle	Item	Difficulty Level	Item/Test Biserial
I*	22	.228	.299
	29	.188	.067
	40	.605	.428
	43	.401	.433
II.1	5	.301	.186
	33	.431	.222
II.2*	1	.702	.131
	35	.593	.428
II.3	3	.425	.354
	49	.588	.475
II.8	17	.628	.409
	47	.493	.476
II.10	26	.449	.538
	36	.494	.403
II.11*	12	.354	.201
	15	.215	.123
III.1	25	.475	.488
	42	.491	.459
III.2*	23	.263	.419
	32	.577	.358
III.3*	39	.314	.350
	48	.633	.448
III.4.e	20	.539	.327
	21	.472	.417
IV.2	27	.469	.292
	31	.466	.349
IV.4	24	.302	.290
	45	.384	.400
IV.8	14	.331	.374
	50	.176	.329
IV.9	38	.456	.211
	41	.333	.333
IV.11*	7	.249	.182
	11	.392	.074
IV.13	9	.276	.150
	13	.233	.282

6.3.4. The Relationship of Item Performance to Overall Test Performance

It is typically assumed that performance on individual items should be substantially positively related to performance on the test as a whole. At the bottom of the assumption is the feeling that if all the items are designed to test "the same thing", then similar results should be achieved across items. Of course, if this "same thing" is a multifaceted thing, the assumption does not hold up. When the domain is multi-dimensional one can imagine situations in which performance on one aspect of the domain is not related at all to performance on some others, or even that the performances are negatively related. However, just because situations like this, though conceptually possible, are difficult to imagine we decided to flag any items whose item biserial correlation was essentially zero or was negative.

Table 6-6 shows that there were no negative item/test biserials and only two (for items 11 and 29) which were essentially zero. Item 11 also had a troublesome thinking/performance relationship and was thus modified. Item 29 had good thinking/performance characteristics but a small change was made to make the first underlined statement less complex, and a note made to see how the item behaved in the subsequent version.

6.3.5. Test Wiseness

Test wiseness is ill-defined. In a rough and ready way we might say that test wiseness affects people's performance on a test if their previous experience taking tests has taught them how to do well on the test through means that are irrelevant to what the test is designed to measure. In addition, this idea can be applied to the internal workings of one particular test if experience taking earlier sections of a test affect in irrelevant ways performance in subsequent parts of the test.

We checked for the effect of test wiseness by examining the protocols and by comparing average performances on different parts of the test. Through the protocols we found that some students chose answers on the basis of the sheer amount that was said by the characters in the story. For example, for item 30 a student chose the first underlined statement (which is the keyed answer) because it "explained more", meaning that it said more. There has been advice in the testing literature for many years to be careful of this type of unwanted influence on test performance, so all questions of the test were examined so that (within particular items) no character in the story gave noticeably more information than another.

We also checked for whether there was any systematic tendency for students to do better in one part of the test or another. Recall that each student took either Part A or Part B of the test and that each of these parts was in turn divided into two sections. Students were interviewed on one section and took the other section in the normal testing manner. Tables 6-9 and 6-10 show how performance scores and thinking scores compared for these groups. Table 6-9 must be read diagonally. For example, cells 1 and 6 contain the average performance scores of a group of students which was first interviewed on Part A, section I, and then took Part A, section II, in the normal testing format. Cells 7 and 4 by contrast contain the average performance scores of a group which first took Part B, section I, in the normal testing format and then was interviewed for Part B, section II. The data show that average performance scores are higher for the second section taken regardless of the part of the test and regardless of whether a group was interviewed before answering questions under normal testing conditions or whether interviewed afterwards. The group represented by cells 7 and 4 is slightly anomalous, since their increase from section I to II is not as great as in the other groups.

On the face of it this increase in performance from section I to II regardless of other factors might be taken as evidence that test wiseness is

influencing performance. However, there is support for the test wiseness hypothesis only when what brings about the increase in scores is irrelevant to what the test is designed to measure. However, Table 6-10 shows that in addition to an increase in average performance scores from section I to II there was an accompanying increase in average thinking scores. This suggests that whatever is the precise cause of the increase, it appears to be quite relevant to what the test is supposed to measure, that is, ability to use principles of thinking to appraise observations. It must be granted that students learned something from doing the test, that this was independent of what, if anything, they learned from the interviews, but that the learning was quite relevant to the purpose of the test. That is the improvement in performance, since accompanied by an improvement in thinking, was likely caused by legitimate reasons and not for something that might be called test wiseness.

One qualification must be added. The average thinking scores in Table 6-10 cannot be directly compared by group to the performance scores in Table 6-9. Table 6-10 contains data on four groups. For example, the group interviewed on Part A, section I, is not the same as the group interviewed on Part A, section II. Thus, without assumptions of initial group equivalence the scores cannot be directly compared. Though the Ns are small, 13 per cell, students were randomly assigned to groups and average performance scores on section I (either Part A or Part B) were essentially the same for all four groups. Thus, there is a strong presumption created of group similarity. The conclusions on test wiseness must, however, be tempered by the possibility of non-equivalence.

Table 6-9: Average Performance Scores by Test Section and by Order of Interviewing

	Part A		Part B	
	Sect. I	Sect. II	Sect. I	Sect. II
Cell Number	1	2	3	4
Interviewed	49	64	47	52
Cell Number	5	6	7	8
Not Interviewed	48	61	48	65

Table 6-10: Average Thinking Scores by Test Section

Part A		Part B	
Section I	Section II	Section I	Section II
34	52	29	32

6.3.6. Misleading Factors, Reading Difficulty, and Clarity of Instructions

This section addresses a number of issues all of which have to do with whether students were able to comprehend the test. To check the overall reading level of the test we used a method which offers suitable accuracy for our purposes and incredible simplicity of application, Fry's Readability Formula.[51] Using the method, the *Test on Appraising Observations* has a

[51] Fry, E. A readability formula that saves time, *Journal of Reading, 11*, 1968, 513-516.

reading difficulty level of between sixth and seventh grade. So, even if Fry's method is considerably inaccurate the test should present no reading difficulties for most senior high school students for whom the test is intended. Despite this low measured reading level, we found during the analysis of interviews that there were vocabulary problems in twelve items, either because students did not know the meaning of particular words or because the meaning was ambiguous. Adjustments were made to correct for each of these difficulties.

We have already discussed problems which arose with the introduction to Version B. However, the instructions also presented difficulties for some students. After reading the instructions in Version B some students were still confused about what to do. We had provided an example item in the instructions but we had shown students only how to mark their answers for the item, not how they might think through it. We thus elaborated the example, showing how someone thinking critically might think about the item and choose and mark an answer.

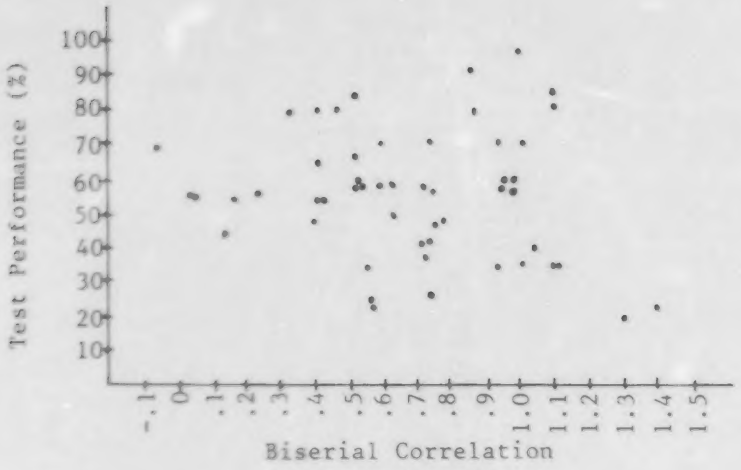
6.3.7. Test Characteristics and Student Ability

We were concerned that some of the test's characteristics might be systematically related to the critical thinking ability of students. In particular, we were concerned that the relationship between thinking scores and performance scores for the test might be dependent upon students' ability. Specifically, if the thinking/performance relationship was related systematically to performance on the test, then we would consider this a significant problem.

Figure 6-1 is a scatterplot of individuals' performances on the test section they took against the biserial correlations of their performance scores and their thinking scores. A visual examination of the plot indicates that there is little systematic relationship, an impression confirmed by a

calculated Pearson's r of 0.12, which is non-significant. There is, then, no systematic relationship between the thinking/performance relationship and test performance.

Figure 6-1: Scatterplot of Test Performance versus Thinking/Performance Biserial Correlation



Theoretically, in the ideal limit there should be a perfect relationship for all individuals between their thinking scores and performance scores. This is the reason for expecting no relationship between performance scores and the thinking/performance relationship: the thinking/performance relationship should be the same, namely unity, regardless of performance. For this reason it is interesting to examine the distribution of thinking/performance relationships across individuals. The correlations range from very slightly negative to unity. Thus, there is substantial deviation from the theoretical limit of all correlations being unity. However, the median of the distribution is 0.70 indicating that a majority of the correlations are substantial. Given a KR-20 reliability of 0.72, this is about the most that could be expected.

6.3.8. Extraneous Influences on Performance Scores and Thinking Scores

For any test it is important to know whether test performance is related to factors like the sex of the examinees and their grade level. In addition, in the test development reported here it is important to know the effect, if any, of such things as being interviewed on the performance and thinking of the examinees. If the interviewing affected people's performance or if the particular person conducting the interviews was a factor, then the value of the interviewing technique as a means of collecting validity information would be diminished.

In order to test whether the interviewing had an effect on test performance an analysis of variance was performed with individuals' performance scores on the four test sections as the dependent variable, and testing format (interview or non-interview), interviewer, test section, grade level, and sex of examinees as the independent variables. Table 6-11 shows that none of the main effects or the two-way interactions were significant. In particular, the interviewing (format) seems not to have affected performance scores, leading us to conclude that the information gathered in the interview was a trustworthy indicator of how the test would work in non-interview situations.

Since performance scores and thinking scores were so highly correlated it was decided to perform a multivariate analysis of variance with thinking scores and performance scores as the dependent variables. The results of this analysis are contained in Table 6-12. There was a grade level effect significant at $>.01$, and a marginal test section effect significant at about .10. These effects are tolerable and represent no invalidating information. In fact, the grade level effect might be expected. There were no significant effects found for the interviewer or for the sex of the examinees. This is a desired result since effects from these sources

Table 6-11: ANOVA Results for Version B: Performance Scores by Testing Format, Interviewer, Test Section, Grade Level, and Sex of Examinee

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Format	1.25	1	1.25	.004	.960
Test Section	3762	3	1254	2.52	.234
Interviewer	63.9	1	63.9	.200	.660
Grade Level	3245	1	3245	5.19	.263
Sex	355	1	355	1.72	.415
For X Int	309	1	309	.947	.334
Int X Ts	1491	3	497	1.52	.215
Int X Gl	626	1	626	1.92	.170
Int X Sex	207	1	207	.632	.429
For X Ts	1811	3	604	1.85	.145
For X Gl	65.4	1	65.4	.200	.656
For X Sex	543	1	543	1.66	.201
Ts X Gl	385	3	128	.393	.758
Ts X Sex	1015	3	338	1.04	.382
Gl X Sex	182	1	182	.559	.457
Within	16410	46	357		

would indicate either that our interview information was untrustworthy or that the test was being influenced by what should be an irrelevant factor.

Table 6-12: MANOVA Results for Version B: Performance Scores and Thinking Scores by Interviewer, Test Section, Grade Level, and Sex of Examinee

Source of Variation	Dependent Variable	Hypothesis Mean Square	Error Mean Square	Significance (Wilks Lamda)
Grade	Performance	2555	191	
				.003
Level	Thinking	195	154	
Test	Performance	.598	191	
				.11
Section	Thinking	577	154	
Inter-	Performance	131	191	
				.549
viewer	Thinking	167	154	
Sex of	Performance	332	191	
				.317
Examinee	Thinking	271	154	

6.3.9. Summary

This concludes the analysis of the data collected on Version B of the test. The aim in all the analyses was to determine the extent to which test performance could be explained by level of critical thinking, and to what extent unwanted factors influenced performance. Our conclusion was that the overall structure of the test was suitable, but that many changes would have to be made at the item and instructions level. We hoped that these local changes would not affect the overall behaviour of the instrument, as

described in the ANOVA and MANOVA results, and in the analysis of test wiseness, reading, clarity, etc.

It has been stressed before that test design is an iterative procedure in which one constructs a prototype and adjusts it until it behaves suitably. The designer is never certain, however, that the upcoming version will be the final one. Thus, we proceeded to construct Version C (see Appendix D) according to the changes indicated thus far and then to collect data similar to that collected for Version B.

6.4. Analysis: Version C

The following questions most concerned us in evaluating the performance of Version C:

1. Did the thinking/performance relationships for items improve over Version B?

2. Did the thinking/performance relationship for the test as a whole improve over Version B?

3. Was performance on items measuring the same principles more uniform than in Version B?

4. Were performance scores as unaffected by unwanted influences as in Version B?

6.4.1. Thinking/Performance Relationships for Items

On Version B there were 16 items with T/P index scores less than .5 and one of these with a negative index score. All of these items but two, items 22 and 23, were altered. We thought that changes made to neighbouring items would favourably affect these two items. Table 6-13 compares T/P index scores per item for Version B and Version C.

Note first that there are no negative T/P index scores for Version C. In addition, the average index score for Version C is .656, an increase

Table 6-13: Comparision of T/P Index Scores Per Item for Version B and Version C

Item	Version B	Version C	Item	Version B	Version C
1	.462	.885	26	.333	.667
2	.731	.885	27	.413	.708
3	.616	.412	28	.750	.708
4	.615	.346	29	.875	.850
5	.885	.731	30	.423	.800
6	.769	.615	31	.346	.800
7	.577	.308	32	.692	.700
8	.385	.385	33	.654	.300
9	.769	.538	34	.538	.750
10	.346	.423	35	.769	.500
11	-.231	.346	36	.461	.450
12	.583	.387	37	.731	.800
13	.458	.654	38	.334	.400
14	.208	.423	39	.833	.550
15	.625	.692	40	.625	.550
16	.536	.625	41	.250	.650
17	.715	.833	42	.375	.250
18	.964	.958	43	.209	.650
19	.786	.917	44	.875	.700
20	.250	.958	45	.667	.500
21	.250	.917	46	.958	.750
22	.429	.750	47	.833	.950
23	.429	.833	48	.917	.800
24	.679	.500	49	.708	.750
25	.769	.958	50	.750	.950
				Mean=.58	Mean=.66
				SD=.219	SD=.205

over the average of .578 for Version B. Of the 15 items on Version B with positive but <.5 index scores, 10 had their T/P index raised to over .5 on Version C. Only one item of the remaining five received a lower T/P index

on Version C. In addition, however, 6 items on Version C obtained index scores $<.5$ which had received scores $>.5$ on Version B. We turned to the protocols collected for Version C and to the patterns of thinking and performance scores to see if any weaknesses in these items could be discovered.

For items 3, 4, 7 and 8 we discovered a common error in students thinking which was largely responsible for the low T/P index scores and which seemed easy to fix. In each case there were two or more students who made their choice on the basis of one of the characters in the story being a driver of an automobile and the other not being a driver. For example, in item 3 three students incorrectly chose the first statement as more believable because Mr. Wang was a driver and would be looking at the traffic closer than a bystander. The obvious fix was to make both characters in each of these four items either a driver or a non-driver, effecting the type of control which is sought in test items.

There remained seven items, numbers 10, 11, 12, 14, 33, 38, and 42, which had T/P index scores of $<.5$. For each of these items most of the negative evidence leading to the low T/P index scores came from a (0,1) combination of thinking/performance scores, that is, from examinees who had answered the item correctly but had received a zero for thinking. For each of these items the (0,1) (T,P) combinations were matched by an equal or greater number of (0,0) combinations. This is a phenomenon which we think is unavoidable by the best test. When examinees do not understand the nature of the problem facing them (as indicated by a 0 thinking score) but they are faced with a multiple-choice format, then their performance scores will often appear as random guesses. That is, for every item that an examinee gets incorrect because his or her thinking was poor, the person will get a proportionate number correct despite thinking poorly. From examining the protocols this appeared to be the mechanism operating for these seven items. The students who thought poorly but got the items

correct appeared simply to be lucky in choosing the correct answer. We were not able to detect anything systematic occurring which led students to pick the correct answers despite poor thinking. That is, the low T/P index scores seemed to reflect more a characteristic of the students with respect to the subject matter in these particular items than it seemed to reflect a characteristic of the items.

6.4.2. Thinking/Performance Relationship for the Test as a Whole

The computed Pearson correlation between thinking scores and performance scores was 0.75. This amounts to an r^2 of .56 and is significant at $>.001$. The corresponding correlation for Version B was smaller, being 0.68. This improvement was judged to be important, and along with the data presented in the previous section, illustrated that the protocols collected on the Version B trial were quite informative in leading to changes which enhanced the validity of the the instrument. In addition, this level of correlation is of the same order as the computed KR-20 reliability coefficient of 0.74.

6.4.3. Items Testing the Same Principle

It was observed for Version B that item difficulty levels and item/test biserial correlations were dissimilar for some pairs of items measuring the same principle. This occurred for 6 principles and involved 14 items. We argued that a certain degree of difference (up to about 50%) is tolerable because even though items might be testing the same principle, the differences in their surrounding context could give rise to legitimate variation. In the end, 10 items were altered to correct for this difficulty.

The changes in the 10 items appropriately affected the characteristics of items testing three of the principles, II.2, II.11, and IV.11. However, for the remaining three principles, I, III.2, and III.3, difficulty levels or

item/test biserials or both continued to differ by more than 50% from each other. In addition, the data for Version C indicated more that 50% differences for three other principles, II.1, III.1, and IV.2. Thus, for Version C there remained 6 principles for which our 50% rule of thumb was violated. On examining each of the items involved, they were found to have suitable characteristics in other regards. For example, the T/P index scores for the items were quite high. Without having any unambiguous information on how to improve the items, we decided to let the items remain as they were. We suspect that there are contextual influences operating in these questions, but without considerable research into the influence of context on people's critical thinking ability we can do little more than speculate, and to produce the best items we can with the knowledge that is currently available.

6.4.4. Extraneous Influences on Performance Scores and Thinking Scores

As for Version B, ANOVA and MANOVA analyses were performed in order to check for unwanted influences on performance and thinking scores. The ANOVA results as shown in Table 6-14 with performance scores as the dependent variable, and testing format (interview or non-interview), interviewer, test section, grade level, and sex of examinees as the independent variables. The analysis showed significant effects for Test Section and Format by Sex interaction. The Test Section effect was of little concern, since there is no specific reason for wanting the difficulty of both sections of the test to be the same. The Format by Sex interaction was, however, of more concern. It did not cast doubt on the validity of the test, necessarily, but it did call into question the applicability of the interview technique in obtaining evidence for validity. However, there was nothing to recommend by the finding in terms of changes to the test, and given that interactions between Sex and other variables were not significant, and given that our main concern that there be no main effect for Sex was satisfied, we merely noted the finding.

Table 6-14: ANOVA Results for Version C: Performance Scores by Testing Format, Interviewer, Test Section, Grade Level, and Sex of Examinee

Source of Variation	Sum of Squares	DF	Mean Square	F	Significance of F
Format	18.0	1	18.0	1.62	.424
Test Section	3615	3	1205	6.84	.074
Grade Level	741	1	741	1.95	.395
Sex	41.8	1	41.8	.165	.754
Interviewer	31.3	1	31.3	.127	.722
For X Int	11.1	1	11.1	.045	.832
Int X Ts	528	3	176	.717	.545
Int X Gl	379	1	379	1.55	.219
Int X Sex	254	1	254	1.03	.313
For X Ts	83.7	3	27.9	.114	.952
For X Gl	180	1	180	.732	.395
For X Sex	1086	1	1086	4.42	.040
Ts X Gl	1511	3	504	2.05	.116
Ts X Sex	190	3	63.4	.258	.855
Gl X Sex	33.2	1	33.2	.135	.714
Within	9320	42	222		

The MANOVA was performed with thinking scores and performance scores as the dependent variables and Grade Level, Test Section, Interviewer, and Sex of Examinees as the independent variables. As with Version B, there was a Grade Level effect, though marginal at about the .11 level. In addition, there was a Test Section effect as with Version B, but with Version C the effect was more highly significant. As hoped, there were no significant Interviewer or Sex of Examinee effects. Results are contained in Table 6-15.

Table 6-15: MANOVA Results for Version C: Performance Scores and Thinking Scores by Interviewer, Test Section, Grade Level, and Sex of Examinee

Source of Variation	Dependent Variable	Hypothesis Mean Square	Error Mean Square	Significance (Wilks Lamda)
Grade Level	Performance	430	126	.108
	Thinking	.282	150	
Test Section	Performance	1494	126	.005
	Thinking	862	150	
Inter-viewer	Performance	7.92	126	.165
	Thinking	351	150	
Sex of Examinee	Performance	15.9	126	.843
	Thinking	7.03	150	

Version C seemed to be a better test than Version B, given the improved T/P index scores and the equally good resistance to extraneous influences. However, it appeared that making improvements from version to version had finally reached a point of diminishing returns. While Version C was better than Version B, it seemed to us that improvements to Version C would not be worth the expense of new data collection. In addition, the data collected on Version C, unlike with previous versions of the test, did not often point unambiguously to changes that should be made. As just discussed, there were some indications that items could be improved, but little information on what sort of change would lead to improvement. Therefore, instead of just "shooting in the dark" we decided to settle for a few cosmetic changes.

To us it seems that we have driven this test construction about as far as it can go without a substantial increase in knowledge about what influences test performances and, in particular, about how critical thinking is influenced. For example, while we have argued that any good critical thinking test must be set in a context because sound appraisals cannot be made devoid of context, there is little knowledge on the effect of context on critical thinking.[52] Thus, given the state of knowledge, we were faced with an unsolvable problem when items purporting to test the same critical thinking principle, but doing so in different contexts, displayed different statistical characteristics. Was this due to differential effects of context on critical thinking performance? If so, it would not point to a problem with the test. Was the difference due to items of varying quality? If so, this would reduce the validity of the test. Until more is known, these questions cannot be answered.

---

[52]Norris, S.P. The choice of standard conditions in defining critical thinking competence. *Educational Theory, 35*, forthcoming.

6.5. Final Data

Data on the final version of the test (see Appendix A) was collected from four high schools in southern Ontario as indicated in Table 6-2. In addition, data was collected on the *Cornell Critical Thinking Test, Level X* and the *Watson-Glaser Critical Thinking Appraisal, Form A*, which were used to compare the characteristics of the *Test on Appraising Observations*. Performance scores only were calculated (the key to correct answers for the *Test on Appraising Observations* is given in Appendix F) as each test was given in normal testing circumstances. An item analysis was performed which yielded item difficulty levels, item/whole-test biserial correlations, and KR-20 reliability estimates. The results for all four schools combined is given in Table 6-16, and for the schools taken separately in Tables 6-17 to 6-20.

Table 6-16: Item Analysis Results for All Southern Ontario Schools Combined

Test	Mean	Variance	KR-20
Test on Appraising Observations	31 (50)*	32	.69
Cornell Critical Thinking Test	51 (76)*	41	.72 (.77-.81)**
Watson-Glaser Critical Thinking Appraisal	49 (80)*	78	.80 (.69-.85)**

* Total number of items on test
** Range of KR-20 reported in test manual
*** Range of split-half reliabilities reported in manual

Table 6-17: Item Analysis Results for Bluevale School

Test	Mean	Variance	KR-20
Test on Appraising Observations	30	41	.76
Cornell Critical Thinking Test	50	42	.74
Watson-Glaser Critical Thinking Appraisal	44	165	.92

Table 6-18: Item Analysis Results for Forest Heights School

Test	Mean	Variance	KR-20
Test on Appraising Observations	31	30	.68
Cornell Critical Thinking Test	51	39	.71
Watson-Glaser Critical Thinking Appraisal	49	54	.72

In terms of average reliability estimates the *Test on Appraising Observations* is the lowest of the three. This is no doubt in part due to the fact that it is the shortest of the three tests, containing 50 items compared to 76 and 80 for the other two tests. The difference also appears partly due to two cases of extreme data in the samples. The variance on the *Test on Appraising Observations* for the Waterloo sample was lower than any

Table 6-19: Item Analysis Results for Guelph School

Test	Mean	Variance	KR-20
Test on Appraising Observations	31	35	.73
Cornell Critical Thinking Test	50	49	.78
Watson-Glaser Critical Thinking Appraisal	51	59	.76

Table 6-20: Item Analysis Results for Waterloo School

Test	Mean	Variance	KR-20
Test on Appraising Observations	32	23	.58
Cornell Critical Thinking Test	51	37	.71
Watson-Glaser Critical Thinking Appraisal	50	52	.74

other measured variance, which contributed to the lowest reliability estimate for that test. By contrast the variance on the *Watson-Glaser Critical Thinking Appraisal* for the Bluevale sample was by far the highest variance recorded, contributing to the high measured reliability. If these two extremes are eliminated from consideration then the spread in

reliability among the three tests would diminish. Overall, we considered the reliability calculated on the entire sample, .69, while low in terms of some psychological instruments, to be adequate given the reliability of the other critical thinking tests.

In addition in terms of some other test characteristics the *Test on Appraising Observations* fared better than the other instruments. Near the end of both the Cornell Test and the Watson-Glaser Test, there was a tendency for items to have zero or negative item/test biserial correlations. This occurred for 7 items on both tests, but did not occur on our test at all. This suggests, but only suggests, that some new factor begins to operate near the end of the other two critical thinking tests. This might be fatigue, due to the length of the test, but this is purely speculation. It would be an interesting issue to pursue, however.

Chapter 7
Summary and Conclusions

This report has described the design of a test of one aspect of critical thinking ability, the ability to correctly appraise observations. The test is intended primarily for an audience of senior high school students, though it might be used effectively at other levels, particularly with undergraduate students. Although the test is intended to be a power test, most senior high school students finish it in one class period, not allowing for administrative time to pass out materials and instruct students in how to take the test and mark their answers. The test can be scored easily by hand, but no machine scoring system is yet in place.

The *Test on Appraising Observations* is based upon a comprehensive set of principles for appraising observations and upon a particular theory of test validation. The principles were described and presented in Chapter 3. Although they are subject to modification, the principles as they currently stand offer the most comprehensive and defensible set known to us. The theory of test validation was presented in Chapter 4 and differs from currently accepted theory in its emphasis on the discovery of the mental mechanisms which lead people to perform on tests the way they do. The theory has given rise to a validation methodology focussed on extracting from examinees their thinking while they work through questions on the test. Chapter 6 describes how this information was subsequently used in assessing the validity of the test.

In the version in Appendix A the *Test on Appraising Observations*

represents an instrument which has evolved through several stages with attendant improvements at each stage transition. The test rates favourably on standard measures of quality with two the the most widely used critical thinking tests, the *Cornell Critical Thinking Test, Level X* and the *Watson-Glaser Critical Thinking Appraisal, Form A*. In addition, this test has been subjected to intensive scrutiny through the protocols of about 100 students. We know that for the most part when students perform well on this test they do so because they have thought well, and when they do poorly they have thought poorly. According to the view of test validation outlined earlier this knowledge is at the foundation of any claim to a test's being valid. We know there are and always will be exceptions. However, with care the instrument can be effectively used. In addition, because the test concentrates on only one aspect of critical thinking ability, it provides a better indication of people's ability in that area than the currently available tests which are designed to give an assessment of critical thinking ability in general.

We have also discovered in this study that protocols of examinees' thinking can be effectively collected and used in designing a test. The protocols were used to identify problem questions, but because we had detailed information on the cause of people's performances we were usually able to turn to the protocols for specific changes that had to be made for items to behave more adequately. The information was very accurate in many cases. Thus, we hope that this report illustrates the effective employment of a validation technique which has a long record of endorsement but only a sparse record of use.

Appendix A

Test on Appraising Observations

test on APPRAISING OBSERVATIONS

This test tells you two stories. Read them very carefully. As you read the stories you will be asked to answer questions about what people say. You must read **ALL** the information you are given. **EACH** piece of information may be needed to answer some questions. Each question has only **ONE** accepted answer.

DO NOT WRITE ON THIS BOOKLET.

STEPHEN P. NORRIS and RUTH KING
INSTITUTE FOR EDUCATIONAL RESEARCH AND DEVELOPMENT
Memorial University of Newfoundland
1983

DIRECTIONS

In each question you will be given two statements in bold type. You must choose which statement in bold type, if either, you have **MORE** reason to believe at the time the statements are made.

Remember: Choose between the statements in bold type only. You may use statements which are not in bold type to help you choose.

Here is an example:

0. Two friends, Cathy and Helen, are driving along a country road. Suddenly an animal runs in front of the car and crosses to the other side of the road.

Cathy says, "Look! **There is a small brown animal!**"

Helen says, "Cathy, you are wearing dark-coloured sunglasses. **That animal was grey.**"

To answer this question, first look for some important difference between the people or the situations. In this case, Cathy is wearing sunglasses. Cathy's sunglasses could have made the animal appear a different colour. Since Helen criticizes Cathy, it seems that she is not wearing sunglasses. Therefore, Helen would have a better view than Cathy. People who have a better view of things tend to be more believable.

Since you have **MORE** reason to believe the **SECOND** statement in bold type, Helen's, at the time the statements are made, you should mark your answer sheet like this:

	First	Second	Neither
0.	O	●	O

Mark your answer sheet now for question 0.

In the rest of the test questions, mark your answers as follows:

First	Second	Neither	
●	O	O	Means you have more reason to believe the **FIRST** statement at the time the statements are made.
O	●	O	Means you have more reason to believe the **SECOND** statement at the time the statements are made.
O	O	●	Means you have no more reason to believe **EITHER** statement at the time the statements are made.

When answering a question, do **NOT** use information given in later questions. You **MAY** use information given in earlier questions. For example, suppose you are working on question 10. You **MAY** use information in questions 1 to 9. You may **NOT** use information in questions 11 to 50.

PART A

A Traffic Accident

A traffic accident has just occurred at an intersection which has a stop sign in each direction. Several cars were involved.

A policeman and a policewoman will question people. Later several investigators will collect information about the accident. It is your job to judge the evidence given in the statements that follow.

1. A policeman is questioning Pierre and Martine. They were in their car at the intersection but were not involved in the accident. Martine is the driver and Pierre, who had been trying to figure out which way to go, is the map reader.

   The policeman asks Martine how many cars were at the intersection when the accident occurred. She answers, "**There were three cars.**"

   Pierre says, "No, **there were five cars.**"

2. A small boy and his father had been standing on the sidewalk when the accident occurred.

   The boy says, "**There was a motorcycle at the intersection.**"

   His father says, "No, **there was no motorcycle at the intersection.**"

3. A policewoman has been asking Mr. Wang and Ms. Vernon questions. She asks Mr. Wang, who was one of the people involved in the accident, whether he had used his signal.

   Mr. Wang answers, "**Yes, I did use my signal.**"

   Ms. Vernon had been driving a car which was not involved in the accident. She tells the officer, "**Mr. Wang did not use his signal.** But this didn't cause the accident."

4. The policewoman then points to Ms. Rosen's car which was one of the cars involved in the accident. She asks whether Ms. Rosen had signalled.

   Mr. Dawe, another driver not involved in the accident, says, "**Ms. Rosen signalled.** I was just talking to Ms. Vernon about this and I'm sure she will agree with what I said."

   Martine says, "**Ms. Rosen did not signal.** I'm sure I'm right."

5. The policeman talks to Mr. and Mrs. Peters, who were also involved in the accident. It is easy to see that Mr. Peters, who was the driver, is very upset by the accident. The policeman asks him to estimate his speed just before the accident.

   Mr. Peters says, "**I was going about 15 kilometers an hour.**"

   A little later when he is feeling well he says, "**I was going about 30 kilometers an hour.**"

6. The policeman asks whether or not the Peters' car had stopped at the stop sign. Ms. Vernon, who is a driver education instructor, says, "I am very experienced in these matters. **The Peters' car did not stop.**"

   Martine, who overheard this conversation, goes up to the officer and says, "**The Peters' car did stop at the stop sign.**"

7. The officer turns to question Martine and Pierre and Mr. Dawe. The officer asks them to estimate the speed of Mr. Wang's car when it hit the others.

   Mr. Dawe says, "**It was going about 40 or 45 kilometers an hour.**"

   The officer says, "It was going faster than that, wasn't it?" Martine says, "Oh yes, **it was going about 60 or 65 kilometers an hour.**"

8. Martine adds, "**Mr. Wang went right through the stop sign.**"

   The police officer turns to Mr. Dawe and says that at the scene of the accident Mr. Dawe couldn't remember whether Mr. Wang had stopped at the stop sign or not. Mr. Dawe says, "I remember now, **Mr. Wang did stop at the stop sign.**"

9. Ms. Vernon then says, "I also remember that **a fancy blue sports car went through the stop sign.**"

   Martine says, "**A car with twin headlights went right through the stop sign.**"

10. Martine says, "Three cars collided at the same time. **There was one crash.**"

    Ms. Vernon says, "**There was more than one crash.** It would be very strange for the three to collide at exactly the same time."

11. The police officers ask the people involved in the accident and the other drivers to come to the police station to make official statements. At the station, the policeman questions Mr. Peters.

    Mr. Peters points to a drawing of the intersection and says, "Just before the accident occurred **Mr. Wang's car approached the intersection from that direction.**"

    The police officer says to Mr. Peters, "Surely Mr. Wang's car came from a different direction." "Oh yes," says Mr. Peters, "**it did come from a different direction.**"

12. The policeman turns to Mr. Dawe to question him. In the background they can hear a conversation between the other officer and some of the other witnesses. Some are discussing whether one of the cars went through a stop sign.

    Mr. Dawe says, "**Mr. Wang and Ms. Rosen crashed into each other.** I saw it happen."

    "Also, I remember that **a car went straight through a stop sign,** too."

13. Nearby, the policewoman and Martine are looking at the drawing of the intersection.

    Martine says, "**A short time before the accident everyone was driving normally.**"

    She continues, "Then there was a loud squeal of tires. **Mr. Peters' car turned quickly toward the fruit stand on the corner.**"

14. The policewoman asks Mr. Dawe to tell in which direction Mr. Peters was travelling before the accident. Mr. Dawe says, "**He was going toward Fifth Street.**"

    The policewoman looks at her notes which were made at the scene of the accident. At that time Ms. Vernon had pointed and said that **Mr. Peters was going away from Fifth Street before the accident.**

15. The policewoman remarks that many people turn left at this intersection even though they are not allowed. She says that this causes many accidents. She asks Martine to continue to tell what she remembers.

    Martine says, "**Ms. Rosen came to a complete stop.**"

    She then adds, "**But then she turned left.**"

16. Meanwhile, at the scene of the accident several inspectors have been collecting information about the accident. They are examining the wrecks and the marks on the road. Two teams were collecting information separately. They are now finished and are comparing notes.

    Inspector Suzuki says, "Our notes say that **Ms. Rosen's car skidded 427 centimeters before hitting the other cars.** I made the measurement and also made the notes."

    Inspector Rousseau says, "According to our notes **Ms. Rosen's car skidded 457 centimeters before hitting the other cars.** Inspector O'Reilly measured the skid by herself and Inspector Smith copied down what she said."

17. Inspector Rousseau says, "We also measured the length of Mr. Wang's skid. We used a measuring tape that was 1000 centimeters long. Inspector O'Reilly held one end at the beginning of the skid and I took the reading at the other end. I wrote down the measurement. **Mr. Wang's car skidded 320 centimeters.**"

    Inspector Rossi says, "I also measured Mr. Wang's skid. I used a 30 centimeter measuring stick. I started by placing one end at the beginning of the skid and by putting a mark at the other end. I then placed the beginning of the stick at that mark, and so on until I reached the end of the skid. I wrote down my measurement. **Mr. Wang's car skidded 350 centimeters.**"

18. Inspector Rousseau says, "I found some brown paint on the left front fender of Mr. Wang's car. I looked at it with a magnifying glass. **It is the same colour as the paint on Ms. Rosen's car.**"

    Inspector Rossi says, "I also studied that paint on the left fender of Mr. Wang's car. I looked at it under the microscope. **It is not the same colour as the paint on Ms. Rosen's car.**"

19. Inspector Smith, who does not use a microscope often, says, "I'd like to check that myself." He looks at the paint sample under the microscope and says, "**There are no gold-coloured spots in this sample.**"

    Inspector O'Reilly, who uses a microscope often, looks at the sample. "**There are gold-coloured spots in the sample,**" she says.

20. Inspector Rousseau and Inspector Smith have been using cameras which develop pictures instantly to take pictures of the accident. Inspector Smith's camera is an older model and is more difficult to adjust. They compare pictures of the skid marks of Ms. Rosen's and Mr. Wang's cars. They are trying to find out who stopped faster.

    Inspector Smith points to his pictures and says, "**Mr. Wang's skid marks are darker than Ms. Rosen's.**"

    Inspector Rousseau looks at his pictures and says, "No, **Mr. Wang's skid marks are no darker than Ms. Rosen's skid marks.**"

21. Both Inspector O'Reilly and Inspector Rousseau have taken pictures of the cars involved in the accident. Inspector O'Reilly says, "My pictures show that **Ms. Rosen's and Mr. Wang's cars were damaged the same amount.** I took several pictures of each car by itself after they were rolled apart."

    Inspector Rousseau says, "My pictures show that **Ms. Rosen's car was damaged more than Mr. Wang's car.** I took several pictures of the pile-up before the cars were rolled apart."

22. Inspector O'Reilly says, "**Mr. and Mrs. Peters' car is only slightly damaged.**"

    She continues, "**The accident probably wasn't Mr. Peters' fault.**"

23. Inspector Rossi and Inspector Suzuki examine the pictures taken by O'Reilly and Rousseau. Inspector Suzuki glances at a picture and says, "**There is a part hanging down under Mr. Wang's car.**"

    Inspector Rossi studies the picture for several seconds and says, "**That's not part of Mr. Wang's car.** That's a shadow."

24. They then turn to examine the wrecked cars. Inspector Rossi points and says, "Look, **the brakeline to the front brakes of Ms. Rosen's car is broken.**"

    Inspector Rousseau overhears this and says, "That's strange. I discovered about an hour ago that **that brakeline was not broken.**"

25. Inspector Smith slides under Ms. Rosen's car to examine the brakeline. "**The handbrake cable is broken,**" he says.

    Inspector Suzuki kneels down and peers under Ms. Rosen's car. "No,", she says, "**the handbrake cable is not broken.**"

26. Inspector Suzuki examines the brakeline of Ms. Rosen's car. She says, "**This rubber hose in the brakeline is worn through.** It must have happened gradually."

    Inspector O'Reilly, who thinks that Inspector Suzuki is always wrong, also examines the brakeline. "No," she says, "**the rubber hose is cut.** It must have snapped suddenly."

27. Inspector Rossi checks the brake fluid container of Ms. Rosen's car. He tells the other inspectors that there is a small amount of fluid left.

    Inspector Smith checks the fluid container as well and says, "**There is no fluid left there.**"

    Inspector Suzuki checks as well and says, "**There is a little left at the bottom.**"

28. Inspector O'Reilly says, "One of the police officers checked the brakes. He told me that he pressed the brakes and **they worked.** Ms. Rosen had at least partial braking power at the time of the accident."

    Inspector Rousseau says, "I just checked the brakes myself. I pressed the brakes and **the pedal went straight to the floor.** Ms. Rosen had no braking power at the time of the accident."